Validating supervised learning approaches to the prediction of disease status in neuroimaging

Alex F. Mendelson

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Medical Physics and Biomedical Engineering

University College London

July 27, 2017

I, Alex F. Mendelson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Alzheimer's disease (AD) is a serious global health problem with growing human and monetary costs. Neuroimaging data offers a rich source of information about pathological changes in the brain related to AD, but its high dimensionality makes it difficult to fully exploit using conventional methods. Automated neuroimage assessment (ANA) uses supervised learning to model the relationships between imaging signatures and measures of disease. ANA methods are assessed on the basis of their predictive performance, which is measured using cross validation (CV). Despite its ubiquity, CV is not always well understood, and there is a lack of guidance as to best practice.

This thesis is concerned with the practice of validation in ANA. It introduces several key challenges and considers potential solutions, including several novel contributions. Part I of this thesis reviews the field and introduces key theoretical concepts related to CV. Part II is concerned with bias due to selective reporting of performance results. It describes an empirical investigation to assess the likely level of this bias in the ANA literature and relative importance of several contributory factors. Mitigation strategies are then discussed. Part III is concerned with the optimal selection of CV strategy with respect to bias, variance and computational cost. Part IV is concerned with the statistical analysis of CV performance results. It discusses the failure of conventional statistical procedures, reviews previous alternative approaches, and demonstrates a new heuristic solution that fares well in preliminary investigations.

Though the focus of this thesis is AD ANA, the issues it addresses are of great importance to all applied machine learning fields where samples are limited and predictive performance is critical.

Contents

List of figures

List of acronyms

ABCV	Approximately balanced cross validation
AD	Alzheimer's disease
ADAS-cog	Alzheimer's Disease Assessment Scale-cognitive subscale
ADHD	Attention deficit hyperactivity disorder
ADNI	Alzheimer's Disease Neuroimaging Intiative (study)
AIBL	Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing
ANA	Automated neuroimaging assessment
ANM	AddNeuroMed (study)
BIBD	Balanced incomplete block design
BICV	Balanced incomplete cross validation
CDR	Clinical dementia rating
CSF	Cerebrospinal fluid
CV	Cross validation
DB-SCV	Distribution balanced stratified cross validation
DOB-SCV	Distribution optimally balanced stratified cross validation
DOF	Degrees of freedom
EKCV	Extended K-fold cross validation
FDG	Fluorodeoxyglucose (18F)
fMRI	Functional magnetic resonance imaging

GIF	Geodesic information flow (algorithm)
GM	Grey Matter
HC	Healthy control (subject)
KCV	K-fold cross validation
KNN	K nearest neighbours
LCM	Lowest common multiple
LDA	Linear discriminant analysis
LOOCV	Leave-one-out cross validation
LPOCV	Leave-p-out cross validation
MCI	Mild cognitive impairment
MMSE	Mini mental state examination
MR	Magnetic resonance
MRI	Magnetic resonance imaging
NB	Naive Bayes
NC	Nearest centroid
NHST	Null hypothesis significance testing
OASIS	Open Access Series of Imaging Studies
PET	Positron emission tomography
QDA	Quadratic discriminant analysis
RBF	Radial basis function
RF	Random forest
RHOCV	Repeated hold-out cross validation
RKCV	Repeated K-fold cross validation
RLT	Repeated learning testing

ROI	Region of interest
RVM	Relevance vector machine
SHOCV	Simple hold-out cross validation
sMRI	Structural magnetic resonance imaging
SPECT	Single-photon emission computed tomography
SPM	Statistical parametric mapping (software)
STEPS	Similarity and truth estimation for propagated segmentations
SUVR	Standardised uptake value ratio
SVM	Support vector machine
TIV	Total intra-cranial volume
VBM	Voxel-based morphometry
WM	White matter

List of common notation choices

Note that some variable names are inevitably reused for different purposes. For instance, l often denotes the number of items available in some set, but it may also be used to signify an arbitrary integer in some sections of the thesis. Where this happens, it should be clear from the context.

$\mathbf{a} = \langle a_i \rangle_{1 \leq i \leq n}$	a sequence of length n
$\mathbb{A}^n$	the set of length n sequences of members of the set $\mathbb{A}$
$\mathbb{A}^+$	the set of arbitrary length sequences of members of the set $\mathbb{A}$
$\mathbb{E}_A[f(A,B)]$	the expectation of $f(A,B)$ with respect to A, where B is held fixed
$\mathbb{V}ar_A[f(A,B)]$	the variance of $f(A,B)$ with respect to A, where B is held fixed
$W = (X,Y)$	an item comprising some features $X \in \mathbb{X}$ and labels $Y \in \mathbb{Y}$
$\mathbb{W} = \mathbb{X} \times \mathbb{Y}$	the space of items
l	the number of total items available
m	the number of items used for training in train-test experiments
n	the number of items used for testing in train-test experiments
E	the number of base experiment repetitions in RKCV
$t \in \mathbb{T}$	a predictor, some function $\mathbb{X} \to \mathbb{Y}$ belonging to the set of predictors $\mathbb{T}$
$u \in \mathbb{U}$	a learner, some function $\mathbb{W}^+ \to \mathbb{T}$ belonging to the set of learners $\mathbb{U}$
$\hat{Y} = t(X)$	the predicted labels of an item
$\phi : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$	a utility metric to evaluate the success or failure of a predictions
$Q = \phi(Y, \hat{Y})$	the real valued performance of prediction $\hat{Y}$
$\bar{Q}$	the mean performance of predictions for many items

$\bar{Q}_k$ the mean performance of predictions for items in the kth testing set

$\mathbf{G} = \langle G_i \rangle_{1 \leq i \leq m}, \mathbf{G} \in \mathbb{W}^+$ a training set, actually a sequence of items

$\mathbf{H} = \langle H_i \rangle_{1 \leq i \leq n}, \mathbf{H} \in \mathbb{W}^+$ a testing set, actually a sequence of items

$\mathbf{D} = \langle D_i \rangle_{1 \leq i \leq l}, \mathbf{D} \in \mathbb{W}^+$ a full sample, a sequence of items that may be divided to make training and testing sets

$I = \{\imath_i\}_{i=1}^m$ a subset of the integers $\{1, 2, ..., l\}$ that may be used to draw a training set of from a full sample

$J = \{\imath'_j\}_{j=1}^n$ another subset that may be used to draw a testing set of from a full sample, typically defined such that $J \cap I = \varnothing$

R the number of train-test experiments used in a train-test experiment

$\mathbf{I} = \langle I_r \rangle_{r=1}^R$ a block design, a sequence of integer subsets that may be used to define the one or more training sets that define a cross validation experiment

$\Gamma(u, \mathbf{D}, \mathbf{I})$ the validation function which returns the mean performance of predictions of the learner u in the cross validation experiment defined by $\mathbf{I}$

M_T the expected performance of a random predictor $T \in \mathbb{T}$

μ_t, μ_t the expected performance of the fixed predictor t or the fixed learner u

Ξ a random variable test statistic

List of publications

1. J. Young, M. Modat, M. J. Cardoso, **A. F. Mendelson**, D. Cash, S. Ourselin, A. D. N. Initiative, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.

2. **A. F. Mendelson**, M. A. Zuluaga, L. Thurfjell, B. F. Hutton, and S. Ourselin. The empirical variance estimator for computer aided diagnosis: Lessons for algorithm validation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 236–243. Springer International Publishing, 2014.

3. M. A. Zuluaga, A. Mendelson, M. J. Cardoso, A. M. Taylor, and S. Ourselin. Multi-atlas based pathological stratification of d-TGA congenital heart disease. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 109–112. IEEE, 2014.

4. M. A. Zuluaga, R. Rodionov, M. Nowell, S. Achhala, G. Zombori, **A. F. Mendelson**, M. J. Cardoso, A. Miserocchi, A. W. McEvoy, J. S. Duncan, et al. Stability, structure and scale: Improvements in multi-modal vessel extraction for SEEG trajectory planning. *International journal of computer assisted radiology and surgery*, 10(8):1227–1237, 2015.

5. N. Burgos, M. J. Cardoso, **A. F. Mendelson**, J. M. Schott, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin. Subject-specific models for the analysis of pathological FDG PET data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 651–658. Springer International Publishing, 2015.

6. A. K. H. Duc, G. Eminowicz, R. Mendes, S.-L. Wong, J. McClelland, M. Modat, M. J. Cardoso, **A. F. Mendelson**, C. Veiga, T. Kadir, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Medical physics*, 42(9):5027–5034, 2015.

7. S. Ferraris, **A. F. Mendelson**, G. Ballesio, and T. Vercauteren. Counting sub-multisets of fixed cardinality. *arXiv preprint arXiv:1511.06142*, 2015.

8. M. Lorenzi, I. J. Simpson, **A. F. Mendelson**, J. M. Cardoso, M. Modat, and S. Ourselin. Voxel-based statistical multimodal model of brain atrophy and hypometabolism in Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(7):P73–P74, 2015.

9. **A. F. Mendelson**, M. A. Zuluaga, B. F. Hutton, and S. Ourselin. Bolstering heuristics for statistical validation of prediction algorithms. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, pages 77–80. IEEE, 2015.

10. J. Young, **A. F. Mendelson**, M. J. Cardoso, M. Modat, J. Ashburner, and S. Ourselin. Improving MRI brain image classification with anatomical regional kernels. In *Machine Learning Meets Medical Imaging*, pages 45–53. Springer International Publishing, 2015.

11. M. A. Zuluaga, N. Burgos, **A. F. Mendelson**, A. M. Taylor, and S. Ourselin. Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries. *Medical image analysis*, 26(1):185–194, 2015.

12. **A. F. Mendelson**, M. A. Zuluaga, B. F. Hutton, and S. Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *arXiv preprint arXiv:1602.05822*, 2016.

13. M. Lorenzi, I. J. Simpson, **A. F. Mendelson**, S. B. Vos, M. J. Cardoso, M. Modat, J. M. Schott, and S. Ourselin. Multimodal image analysis in Alzheimers disease via statistical modelling of non-local intensity correlations. *Scientific reports*, 6, 2016.

14. **A. F. Mendelson**, M. A. Zuluaga, M. Lorenzi, B. F. Hutton, and S. Ourselin. Selection bias in the reported performances of AD classification pipelines *NeuroImage: Clinical*, 14:4–416, 2017.

Chapter 1

Introduction

As the world's population ages, a growing fraction lives with ageing related cognitive impairment [1]. When this cognitive impairment becomes severe enough to interfere with a person's work and usual activities, it may be called dementia [2]. There are multiple pathological processes responsible for cognitive decline. Of these, Alzheimer's disease (AD) is probably the most common [3], followed by vascular dementia. As of 2016, about 40 million people have dementia worldwide, and this number is expected to increase to 100 million by 2050 [1, 4].

AD and other dementias have a heavy social cost, as they greatly reduce their sufferers' quality of life [5]. They also place a great financial burden on individuals and governments responsible for sufferers' long term care [5]. The total monetary cost of dementia was estimated at 604 billion USD in 2010 [6]. For AD, as for most other causes of dementia, there is currently no cure.

The great societal burden of dementia has prompted much research into new methods to improve the measurement and tracking of its underlying pathologies. Clinical trials for new therapies rely on accurate measurements of response to treatment, and preventative therapies are only effective when pathology can be detected early on. Neuroimaging offers an unparallelled description of the brain's structure and physiology, but the information it provides is not easy to interpret. For these reasons, many new computational methods have been developed to better extract meaningful clinical and physical quantities from neurological images. These include the methods of automated neuroimaging assessment (ANA) and diagnosis, which use machine learning to directly estimate the level or category of pathology in a person.

A great deal of research effort is expended to refine and develop new ANA pipelines in search of improved performance, particularly for applications in AD [7]. However, without reliable ways to validate and compare these pipelines, many of the apparent gains will be illusory, and much of this research effort will be wasted. A thorough and considered review of validation practices therefore has the potential to be of great benefit to the field.

1.1 Biomarkers for AD

AD is a progressive condition whose symptoms are initially mild. The brain atrophy and tissue destruction associated with it are likely to be irreversible [8], making a true cure unlikely. For this reason, clinical trials have focussed on finding therapies to slow and prevent tissue loss in an early phase of the disease before large scale changes can occur. Clinical trials for new therapies depend on both accurate tracking of disease progression and a reliable pre-selection of subjects who are most likely to benefit. Pre-selection will typically involve identifying incipient AD in a cohort of subjects suffering from mild cognitive impairment (MCI), a heterogeneous condition with a number of other potential aetiologies [9, 10]. Tracking will involve monitoring differences in primary and secondary outcome measures between treatment and control groups.

The term biomarker refers to a reliable measure of disease progression of the type that may be used for pre-selection and outcome measurement in a clinical trial. More effective biomarkers with lower natural variability allow for clinical trials to be more powerful, smaller, and shorter. In a world where inconclusive results may go unreported [11], higher power is also crucial in ensuring the reliability of published results [12]. Smaller trials place fewer patients at risk of side effects. Because the patent for a new drug lasts a fixed term that begins with the drug's discovery, longer clinical trials provide a financial disincentive for commercial investment in research [13]. The identification of the superior biomarkers for both of these tasks is thus an important research goal in itself [14]; for AD, as for cancer and heart disease, "research investments aimed at establishing and validating surrogate endpoints may have a large social return" [13].

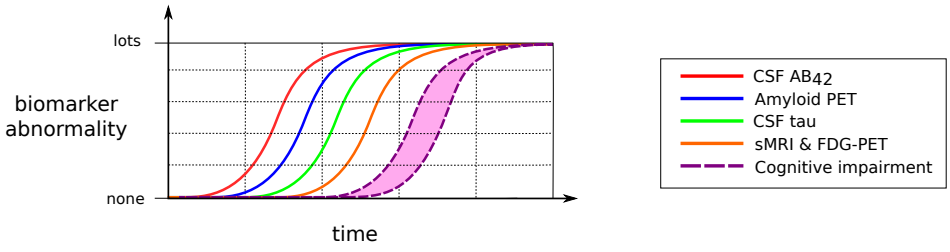

Figure 1.1: The sequential biomarker change model of [15]. The two dashed lines describing the progression of cognitive impairment denote the variability of onset time and decline rate.

A variety of biomarkers have been proposed for pre-selection and response tracking in AD [14]. These include measures derived from biopsies, imaging, and cognitive tests. Those derived from imaging may be called imaging biomarkers. Different biomarkers show their most dramatic changes at different stages in the progress of the disease, with cognitive and behavioural change being the last. Much of the research currently underway is guided by the

theoretical model described in [15]. In this model, the first marker of AD is the formation of amyloid plaques and neurofibrillary tangles (NFTs) in the brain. This can only be measured precisely through post-mortem histology, though it can be less accurately inferred through changes in the protein composition of the cerebrospinal fluid (CSF) [9, 15]. A biopsy of the CSF in an AD patient will reveal elevated levels of tau proteins responsible for NFTs. It will also reveal reduced levels of Aβ_{42}, a particular amyloid protein fragment, as it is accumulated rather than cleared from the brain [16]. Shortly after amyloid deposition is revealed in the CSF, it may also be detected with positron emission tomography (PET) using a family of radiotracers that bind to fibrillar amyloid [10, 17]. After amyloid deposition, the next biomarkers to visibly change are those related to brain structure, as measured using structural magnetic resonance imaging (sMRI), and glucose metabolism, as measured using PET with a ^{18}F fluorodeoxyglucose (FDG) radiotracer [15]. This sequence of biomarker changes is presented in figure 1.1, and an illustration of the associated imaging signatures is presented in figure 1.2.

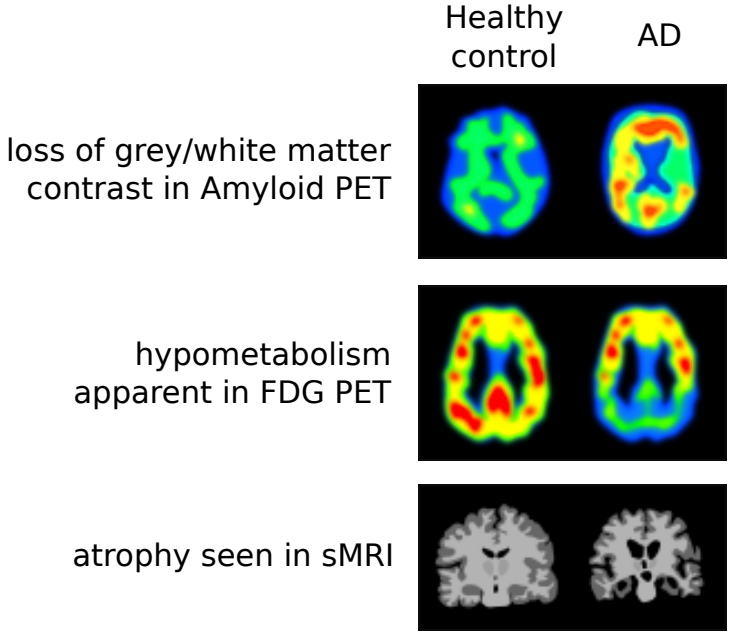

Figure 1.2: Examples of medical image pairs showing signatures typical of healthy subjects and patients with advanced AD. Above, increased levels of fibrillar amyloid in AD reduce the grey/white matter contrast seen in amyloid PET. Centrally, the hypometabolism associated with the disease is apparent in FDG-PET. Below, the widespread tissue atrophy associated with late stage AD is seen in sMRI.

1.1.1 Imaging biomarkers

Imaging data take the form of a large number of voxels. This information must be reduced to some useful univariate quantity before it can be used as biomarker. For a long time, this has been done using visual rating scales such as the Scheltens medial temporal lobe atrophy

score for sMRI [18]. While visual scales are easily translated into routine clinical practice, they lack precision and suffer from inconsistencies between raters [19, 20]. Region-based quantitative methods such as hippocampal volumetry in sMRI and reference-normalised standardised uptake value (SUVR) in PET [21] can offer biomarkers with greater statistical and diagnostic power [19]. Where these methods are based on manual delineation of regions of interest (ROI), they may be time consuming to produce and, to an even greater extent than visual rating scales, dependent on rare clinical expertise [19]. Where different human raters are associated with different groups of subjects in a clinical trial, this leads to systematic differences that can cause spurious effects [20]. Where regional measurements can be automated, these restrictions on their use are eliminated.

Though they can offer precision and repeatability, region-based measures necessarily discard potentially useful information from the majority of voxels. Region-based biomarkers must also be '*hand-crafted*' for each new application; that is, they are dependent on the selection of a meaningful region-based on results in earlier studies. When the progression of disease may actually entail subtle changes distributed across the entire brain, region-based methods may be overly reductive [22]. Even if an optimal region selection could be guaranteed, region-based imaging biomarkers may still fail to capture more complex patterns of change; not all relevant regions may be equally informative, and there may be valuable information in their joint distribution. In the last decade, advances in machine learning have provided automatic methods that can be used to overcome these limitations, and effectively perform a data-driven determination of the relevant regions and their appropriate weightings [23].

1.2 Machine learning

Machine learning can be viewed simply as the application of statistical analysis with a practical focus; its techniques offer automatic ways to produce hypotheses from example learning data. All machine learning methods are inherently statistical, not necessarily in the sense that they involve generative models, but in the sense that they consider the training data to be the result of a random process. Classical statistical techniques are often concerned with explaining or interpreting observed events to allow humans to make well informed decisions. Machine learning techniques more often try to make those decisions directly; the predictive models they produce may not be of direct interest themselves, but they are wanted mainly because they offer high performance in some predictive or decision making task.

Machine learning may be itself divided into multiple overlapping subfields. In the last 15 years, the following division is perhaps the most common choice:

unsupervised learning, which is concerned with the discovery of structure in unlabelled data and includes methods for clustering and dimensionality reduction;

supervised learning, which concerns learning a function which maps inputs to outputs and includes methods for regression and classification; and

reinforcement learning, which concerns agents taking actions in an environment to maximise cumulative reward.

Of these, supervised learning is perhaps the dominant subfield, with the terms supervised learning and machine learning sometimes being used interchangeably. The term pattern recognition often signifies machine learning in the context of machine vision. Supervised learning, and particularly classification, will be the main topic of this thesis. A more formal description of supervised learning will be presented in chapter 3, but a brief outline is also included here.

1.2.1 Outline of supervised learning

In supervised learning, the data comprise a series of atomic observations or **items**. Each item $Z = (X, Y)$ is an ordered pair of two variables: some descriptive **features** that are always available (denoted $X \in \mathbb{X}$), and some dependent **labels** that may be either available or hidden (denoted $Y \in \mathbb{Y}$). Where X may be represented as a sequence of d numeric values, d is termed the **dimension** of the feature space. In order to predict the labels when they are hidden, one must use some **predictor** $t : \mathbb{X} \to \mathbb{Y}$ belonging to the set of predictors $\mathbb{T}$. The purpose of a supervised learning method or **learner** is to select the predictor T based on a **training set** of labelled items. The term **classification** refers to problems where the labels are categorical, while the term **regression** refers to problems where the labels are ordinal or real valued. In classification, each possible label value is referred to as a **class**.

1.3 Automated neuroimaging assessment and diagnosis

Automated neuroimaging assessment (ANA) is defined as the use of supervised learning methods to infer some clinical variable describing the severity or type of disease present in a person based on neuroimaging data. At the cost of some interpretability, supervised learning methods may offer biomarkers with improved sensitivity to change and diagnostic performance [22, 24]. In ANA, items are people, features are imaging descriptors, and labels are relevant clinical variables. I have chosen the term 'automated neuroimaging assessment' over 'computer aided diagnosis' [25], as that term has been used to describe any non-trivial use of a computer output in a diagnostic context [26], and the term diagnosis may be taken to exclude methods that attempt to infer only disease severity rather than type.

The particular benefits offered by ANA will depend on the application: as discussed in section 1.1, imaging biomarkers able to offer unparallelled pre-selection and response tracking can allow for increased power and reduced sample sizes. Even if imaging biomarkers are only comparable in power to existing measures, they may be desirable if they are less invasive (e.g, than biopsy) [22]. The benefits of a diagnostic method that can surpass that of current expert radiologists are obvious, but even those with comparable performance may have great utility; in a clinical setting, automatic methods may provide a rapid alternative to an expert radiologist who may be either unavailable or slow to respond [27]. As well as predicting the presence or absence of disease, supervised methods may be trained to predict response to treatment. By so doing, they may prevent the needless infliction of side effects on patients who are unlikely to benefit [27]. In some cases, the prediction models constructed by supervised learning methods may themselves be used to study the imaging footprint of a condition, though this is typically a secondary goal.

The use of supervised learning to build predictive models requires a training set of labelled examples as input. This dataset comprises a set of pairs of neurological images and corresponding disease states. This disease state may be the diagnosis of an expert physician [25], a psychological score [22], a biopsy measure [22], or an outcome derived from follow-up [28]. In the last case, the prediction to be made is not about the status of a person at the present time, but at some point in the future. A typical pipeline begins with some image processing step, which may involve registration, tissue segmentation, and delineation of relevant anatomical regions. This is followed by some feature extraction, where the resultant images are converted into some appropriate descriptors called features. In some cases, a dimensionality reduction or feature selection technique may then be applied to make the number of features more manageable. Finally, some supervised learning technique is applied to build a predictive model linking the feature description of an image and the corresponding disease state. This model can then be applied to the images of unseen subjects to infer their disease states [23–25, 27]. An illustration of this paradigm is presented in figure 1.3.

This thesis will focus on ANA applications in AD. While dementia, and specifically AD, is probably the most studied application in ANA research [24, 27, 29], it is only one of many. ANA has also been applied to problems in schizophrenia [30], depression [31], attention deficit hyperactivity disorder (ADHD), and many other neurological disorders [32]. All these fields share a strong focus on the development of new imaging features and novel algorithms with the aim of improving performance [22] and a set of common validation challenges.

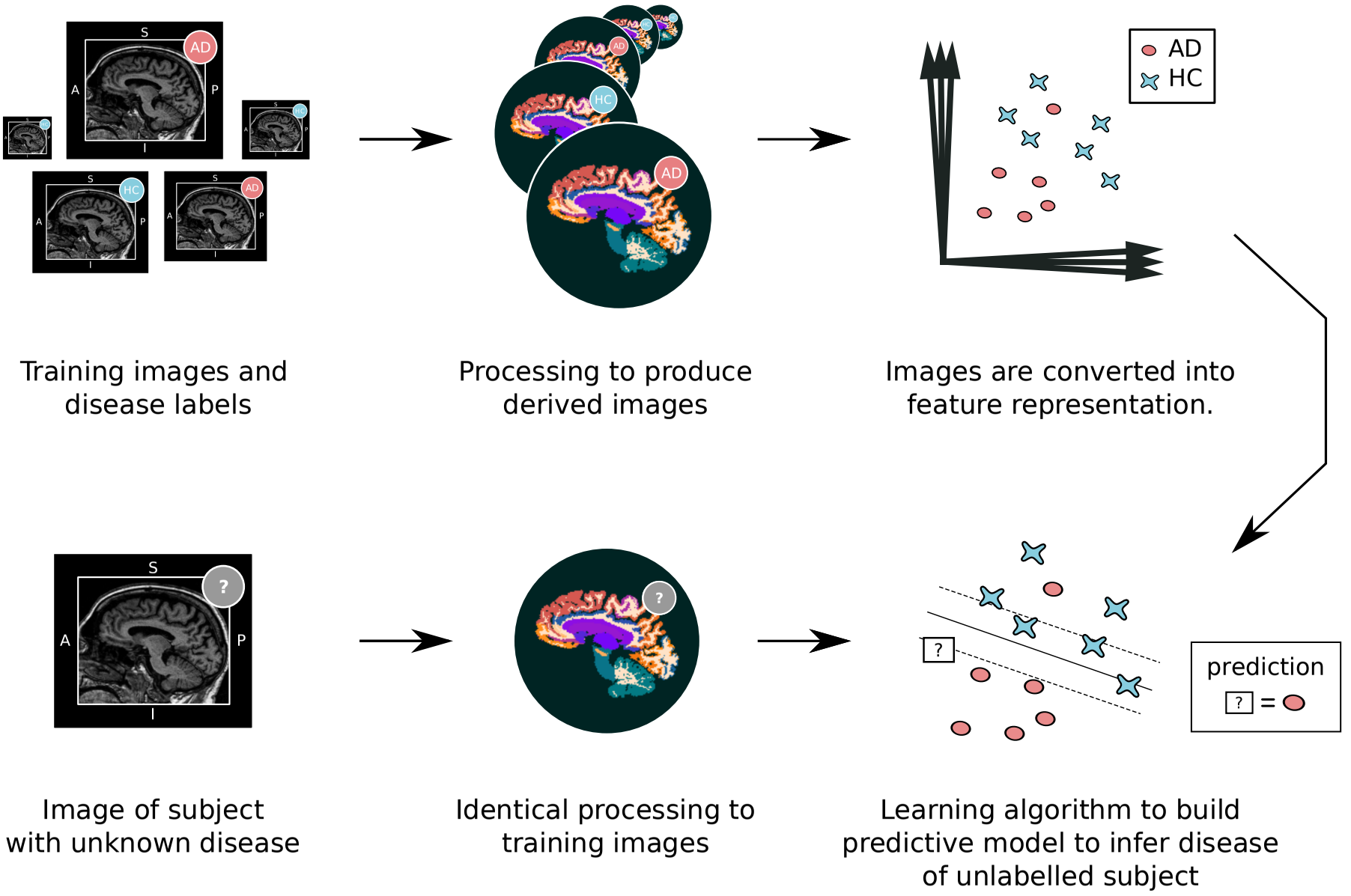

Figure 1.3: Illustration of the composition of a typical ANA pipeline. In this case, the disease description is categorical, being either AD or healthy control (HC).

1.4 Validation challenges in ANA

AD ANA research is a field characterised by a large number of researchers [25] working with necessarily limited data [29] to search for methods with superior performance. The collections of imaging and clinical data required to validate ANA pipelines are very expensive; it is not easy to produce them, and they are unlikely to ever reach the size of the training datasets used in more general applications of machine vision [33]. After nearly a decade of pipeline development [34], any clinical trial or healthcare provider wishing to apply an automatic method now has a large number of options to choose from [24, 25]. There is no clear best option, and there are a variety of issues that can make it difficult to generalise published results [27, 29, 35]. For all the effort expended to develop new methods, surprisingly little has been invested in identifying the most appropriate validation strategies. Without reliable and convincing statements about the relative performance of new methods, the development of those methods is of little practical use.

This section describes the four key validation challenges facing ANA researchers that I have identified.

1.4.1 Statistical analysis of performance results

The use of supervised learning in ANA necessitates some form of cross validation (CV), broadly defined as the use of separate training and testing sets, to estimate the performance of methods [36]. Unfortunately, the component performance measurements of CV are not independent [35, 37, 38]. The classical statistical techniques that are used to quantify uncertainty (hypothesis tests, confidence intervals, etc.) rely on assumptions of independence between observations. When they are used in the context of CV, they may no longer provide the securities that justify them; confidence intervals may have coverage below the nominal level, and the type I error rates of hypothesis tests may be inflated [35]. That is, 95% confidence intervals may not contain the true value 95% of the time, and p values less than 5% may occur with probabilities of more than 5% under the null hypothesis. In this thesis, I shall call this issue the **problem of dependency**.

This problem is particularly relevant to applied fields of machine learning where the focus is on a single prediction task, and the amount of available data is small (as is typically the case in ANA). When researchers are concerned with the performance of a method in a general context, they can evaluate it on multiple independent samples corresponding to different representative problems; the performance measurements from each problem are then independent, and classical statistical procedures can be used [39]. When samples are large enough, the distinction between a learning algorithm and the model it generates can be neglected, as model parameters

will become approximately constant. In such a context, the results on the items of a test set may be considered independent [40].

In ANA applications, where the problem of dependency is not acknowledged, statistical assessments of performance may be unreliable. Where it is, researchers may omit statistical treatment of uncertainties entirely for fear of being criticised. This situation is arguably even worse, as conclusions made on the basis of point estimates alone will be even less reliable than those made using inappropriate techniques. Without reliable treatment of uncertainties, there is little guarantee that published results will generalise to practical contexts. If results do not generalise, they are not useful.

1.4.2 Selection bias

The second validation problem facing ANA methods researchers is one seen in many other fields: publication bias [12, 41]. An inevitable consequence of the search for superior methods is that experiments showing high performance results are more interesting than those that do not, and the latter are more likely to go unreported. The presence of random effects in the measurement of performance means that impressive results can occur by chance. When researchers individually or collectively measure the performances of a wide set of pipelines and then report only the more impressive results, a large number of the reported measurements are likely to be those where the pipeline was "lucky" and had an unusually good result. When a high performing pipeline from the literature is applied to independent data, it is likely to have a performance that is worse than the reported estimate. This effect is due it being unrepresentatively well suited to the particular testing items on which it was first evaluated, rather than unusually well suited to the particular *population* of those items. In the context of AD, automatic methods for pre-selection and differential diagnosis may fail to provide the performance demonstrated in published research, even if study conditions are perfectly replicated. This failure of reproducibility has the potential to undermine the credibility of the ANA research field.

1.4.3 Variance

Small sample sizes are related to higher variance in the estimation of performance. This variance is a problem in itself, as it limits precision. In the context of pipeline refinement, improvements in performance that are smaller than the natural variation in a validation experiment will not be consistently detected. Variability is also intimately related to selection bias, with larger variances producing greater publication bias [12, 41]. It is therefore of great interest to identify efficient validation strategies that are able to provide lower variance using a sample of a given size.

1.4.4 Population drift

The bias discussed in section 1.4.2 results from pipelines being unrepresentatively well tailored to a *particular sample*. Another source of bias, called population drift, results from pipelines being unrepresentatively well tailored to a *particular population* [42]. Though it is not dealt with in this thesis, it must be mentioned for the sake of completeness, as it is likely to be very important in any translational context. If ANA methods are ever to be brought into the clinic, they will have to contend with heterogeneous patient populations that are very different from those seen in studies, and scanning technology that is likely to be of lower quality [27,43]. The extent of this effect was measured in several contexts by the authors of [22], who found that this effect tends to introduce an optimistic bias. To some extent, this problem can be overcome by including data from multiple studies, centres and populations in the datasets used for ANA development and validation.

1.5 Original contributions

This thesis contains novel contributions in the following three areas.

Selection bias. I review the problem of selection bias in ANA, and design an experiment to measure this bias empirically. I use this experiment to demonstrate that bias can account for a significant fraction of the apparent improvements associated with pipeline AD classification optimisation in finite samples. I am able to identify the key factors responsible for bias, and point towards better validation practices that may be used to reduce bias and detect it where it has occurred.

Variance. I discuss the merits of different CV strategies for ANA and describe a trivial extension to repeated K-fold CV that allows for greater experimental flexibility. I also develop and validate another, more complicated extension that has lower variance while using the same amount of computational effort.

Statistical analysis of performance results. I develop and validate a new approach for the construction of heuristic statistical procedures for CV results. The new statistical procedures are shown to have better power and lower type I error/higher interval coverage than conventional alternatives.

1.6 Outline

The remainder of this thesis is structured as follows. Part I contains detailed background information, with chapter 2 providing a detailed review of ANA for AD, and chapter 3 providing

a formalisation of the supervised learning problem. Part II deals with bias due to the selective reporting of performance results, with chapter 4 offering the relevant background, and chapter 5 describing an empirical investigation to estimate the degree of selection bias present in the field and identify best practices for its reduction. Part III deals with the optimal selection of CV strategy. Chapter 6 reviews uncommon CV strategies from the literature and discusses the issue of strategy parameter selection. Chapter 7 describes my work to develop a new strategy with greater efficiency and parameter flexibility than existing alternatives. Part IV deals with the statistical analysis of CV performance results, with chapter 8 detailing the problem of dependency between component results, chapter 9 reviewing proposed statistical tests from the literature, chapter 10 discussing the design and validation of a heuristic rule for the construction of new statistical tests. Finally, part V concludes this thesis by discussing the implications of the work presented for the field of AD ANA research and the possible directions in which it might be extended.

Part I

Background

Chapter 2

Automated neuroimaging assessment for Alzheimer's disease

In this chapter, I shall review the key materials and methods of AD ANA research. The methods selected for discussion are those that appear later in this thesis and those that are important for the field of ANA in general. I shall begin by describing learning tasks in section 2.1, and then move on the imaging modalities used to provide imaging features in section 2.2. In section 2.3, I shall describe the datasets which provide imaging data. In section 2.4, I shall describe the crucial imaging tools necessary to produce meaningful imaging features. I shall then describe what these are and how they are processed in section 2.5. Finally, in section 2.6, I shall describe some important supervised learning algorithms.

2.1 AD ANA Tasks

ANA methods are studied for their potential use as biomarkers. As discussed in section 1.1, biomarkers for AD are primarily intended for use in the following tasks: 1. the identification of incipient AD in its earliest phases for clinical trial enrichment and early intervention and 2. the provision of a reliable measure of disease progression that can be used to track therapeutic response in clinical trials. In practice, a variety of surrogate tasks are used to evaluate possible pipelines, with the most common being classification based [29]. These include the following, with the first three being the most common:

1. Discrimination of AD patients from healthy controls [34, 44]. This can be viewed as a learning exercise for the others that follow; if a method cannot identify AD in its later stages, it is unlikely to be able to do so in its earlier stages. The detection of advanced AD using imaging has no clinical utility in itself, as this can be done more easily through basic questionnaires such as the mini mental state examination (MMSE).

2. Discrimination of MCI subjects and healthy controls [45, 46]. This is a harder task, but

still one of relatively limited utility in itself.

3. Prediction of progression from MCI to AD in a given time window after data collection [28, 47]. This task is harder than the others, but it is the most representative of the intended application of an ANA method. Its difficulty depends heavily on the choice of the time window considered, typically of the order of a year. When the time window is smaller, the progression events to be predicted will be more imminent. The progressive subjects to be identified will then be those at a more advanced stage of AD, making their identification an easier task.

4. Differentiation of AD and other dementias such as frontero-temporal lobe degeneration [48, 49], a task that is also relevant for the enrichment of clinical trials.

Less commonly, regression surrogate tasks are used, such as one of the following:

- Estimation of psychological scores such as the ADAS-Cog and the MMSE [50, 51]. The fine "resolution" of these indicators of disease progression (relative to diagnostic labels) may provide more information in the training of predictive models, though it may come with more "noise".

- Prediction of the time to progression from MCI to AD [52, 53]. This task is similar to the prediction of progression in the classification equivalent.

2.2 Imaging modalities and other information sources

By far the most common choice of imaging modality is MRI, due to its ubiquity and lack of ionising radiation. In particular, structural magnetic resonance imaging (sMRI) is well established as a way to measure the loss of neural tissue associated with AD, and so it is a natural choice of information source [22, 24, 27]. Alternatively, functional magnetic resonance imaging (fMRI) may be used to measure the disruption of the brain's functional networks [54, 55]. More rarely, diffusion tensor MRI may be used to exploit changes in microstructure that may precede the larger scale changes apparent in sMRI [56]. FDG-PET, which can detect changes in cerebral metabolism, has been extensively studied for AD ANA purposes [57, 58]. Amyloid PET has been used too [59], though to a much lesser extent. Even single-photon emission computed tomography (SPECT) has been used [60, 61] for its ability to measure changes in cerebral blood perfusion.

Commonly, only a single image is used to describe each subject in the sample, but it is also possible to use several images from multiple time-points [62, 63]. This allows one to

see not only the morphology/physiology in a subject, but also to see its rate of change. While longitudinal data provide more information, scans at multiple time-points are likely to be harder to achieve in a clinical or practical setting. Most longitudinal methods are based on MRI, whose ubiquity makes this more feasible, though PET data have also been used [57].

The modalities mentioned above may be used alone or in combination. By incorporating information from complementary sources, multi-modal methods aim to achieve greater performance than using any one alone. Perhaps the most common combination is sMRI and FDG-PET [28, 45, 50], though various other combinations have been demonstrated [29, 64, 65]. In addition to combinations of multiple imaging modalities, various combinations of imaging and non-imaging information have been explored. Non-imaging data sources considered include measures based on genetics [28, 66], blood composition [67, 68], CSF biopsies [28, 69] and psychological tests [70].

2.3 Datasets

In order to train and validate new methods, ANA research requires large collections of standardised clinical and imaging data. Due to the necessary expense, it would never be possible for individual studies to gather their own data. As such, the field is reliant on large data sharing initiatives created for biomarker research. These include the following.

The Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI[1] is a large multi-centre study with 1000's of subjects across the U.S. and Canada. It has been developed to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials [14].

AddNeuroMed (ANM). ANM is a multi-centre European study with 100's of subjects. It aims to develop and validate novel surrogate markers of disease and treatment, based upon *in vitro* and *in vivo* models in animals and humans in AD. The neuroimaging part of ANM uses MRI and magnetic resonance spectroscopy to establish imaging markers for early diagnosis and detection of disease and efficacy of disease modifying therapy in man, as well as translational imaging biomarkers in animal models of AD. It has been designed for compatibility with ADNI, and uses the same MRI protocols [71].

[1] www.adni-info.com

The Australian Imaging Biomarkers and Lifestyle (AIBL). AIBL is a two-centre Australian study comprising over 1000 subjects. It aims to discover which biomarkers, cognitive characteristics, and health and lifestyle factors determine subsequent development of symptomatic AD [72].

The Open Access Series of Imaging Studies (OASIS). OASIS is a series of magnetic resonance imaging datasets that is publicly available for study and analysis. The dataset consists of a cross-sectional collection of 416 subjects aged 18 to 96 years. One hundred of the included subjects older than 60 years have been clinically diagnosed with very mild to moderate AD [73].

Of these, ADNI appears to provide by far the largest samples, and it also is by far the most commonly used [24, 25, 29].

2.4 Imaging tools

ANA pipelines rely on meaningful imaging features capable of describing the changes associated with disease. This section describes the key imaging tools needed for their extraction.

2.4.1 Registration

Image registration is the process of deforming images to ensure local correspondence. In medical image registration, anatomical images are deformed to ensure that, after deformation, a given pixel or voxel has a consistent anatomical interpretation. In ANA, this is necessary to make the comparison of voxel intensities meaningful. Typically, image registration is pairwise, and is defined based on the registration of one floating image to the space of another target image. In this case, the floating image is deformed and resampled to the size of the target in a way that minimises some notion of difference. Each different target to which an image can be registered has its own coordinate system or space. In the last 20 years, a large number of registration methods have been developed [74], and many of these have been made publicly available online.

The crucial ingredients of an image registration technique are the following:

- a cost function to quantify a notion of difference between images,
- a parametrisation of the allowed set of transformations, and
- an optimisation technique to search through the space of allowed transformations.

Of these, the ideal choice of cost function will depend on the relationship of the images to be registered. When they are of the same modality, one may simply used the squared sum

of differences in voxel intensities. In other cases, where the intensity profiles of the images do not match, one may use cost functions based on mutual information. The parametrisation used determines the level of flexibility of the transformation. For intra-subject registration, a **rigid** transformation will be most appropriate. For intra-subject registration using images with low anatomical resolutions, such as those from PET or SPECT, one may wish to allow shearing and scaling as well as rotation. The appropriate family of transformations in this case is the **affine** one. Finally, when registering images with high anatomical resolution, such as those of sMRI, one may use non-linear registration. This may be parametrised using a grid of control points with interpolant splines [74, 75] or through the movement of an elastic or fluid substance [76]. Non-linear registration methods typically incorporate regularisation terms into the cost function to prevent infeasible, highly convoluted transformations. An illustration of the different transformation parametrisations is provided in figure 2.1.

2.4.1.1 Groupwise registration

Image registration is based on the **pairwise registration** of floating-target image pairs, but it is often necessary to align a sample of more than two images in a process often termed spatial normalisation. One simple approach is to pick a single target image and to register all images in the sample to that. This target may be chosen from the sample, or may be a pre-constructed template image built from one or more exemplary images. However, because deforming and resampling image causes some degree of degradation that increases with the degree of deformation, this may not be a good choice. A target that is dissimilar to the images of the sample will result in unnecessary degradation, and may introduce unwanted systematic effects when one part of the sample is more similar to it than another. To avoid this, one may use **groupwise registration**, which constructs a representative template target from the sample itself. To do this, one selects an initial target, registers all images of the sample to it, and then combines them by averaging to produce a target that is more representative of the sample as whole [77]. The target image produced this way is initially blurred and poorly defined, but when the registration and combination steps are repeated multiple times, it becomes more well defined. The registration used for the initial steps should initially be based on a highly constrained parametrisation (e.g., affine), which may then be relaxed in later iterations to allow for more accurate anatomical correspondence (e.g., non-linear).

2.4.2 Atlas parcellation

It is often necessary to partition the brain into various anatomical regions of interest. This allows one to consider, for example, the amount of grey matter in a relevant region of the

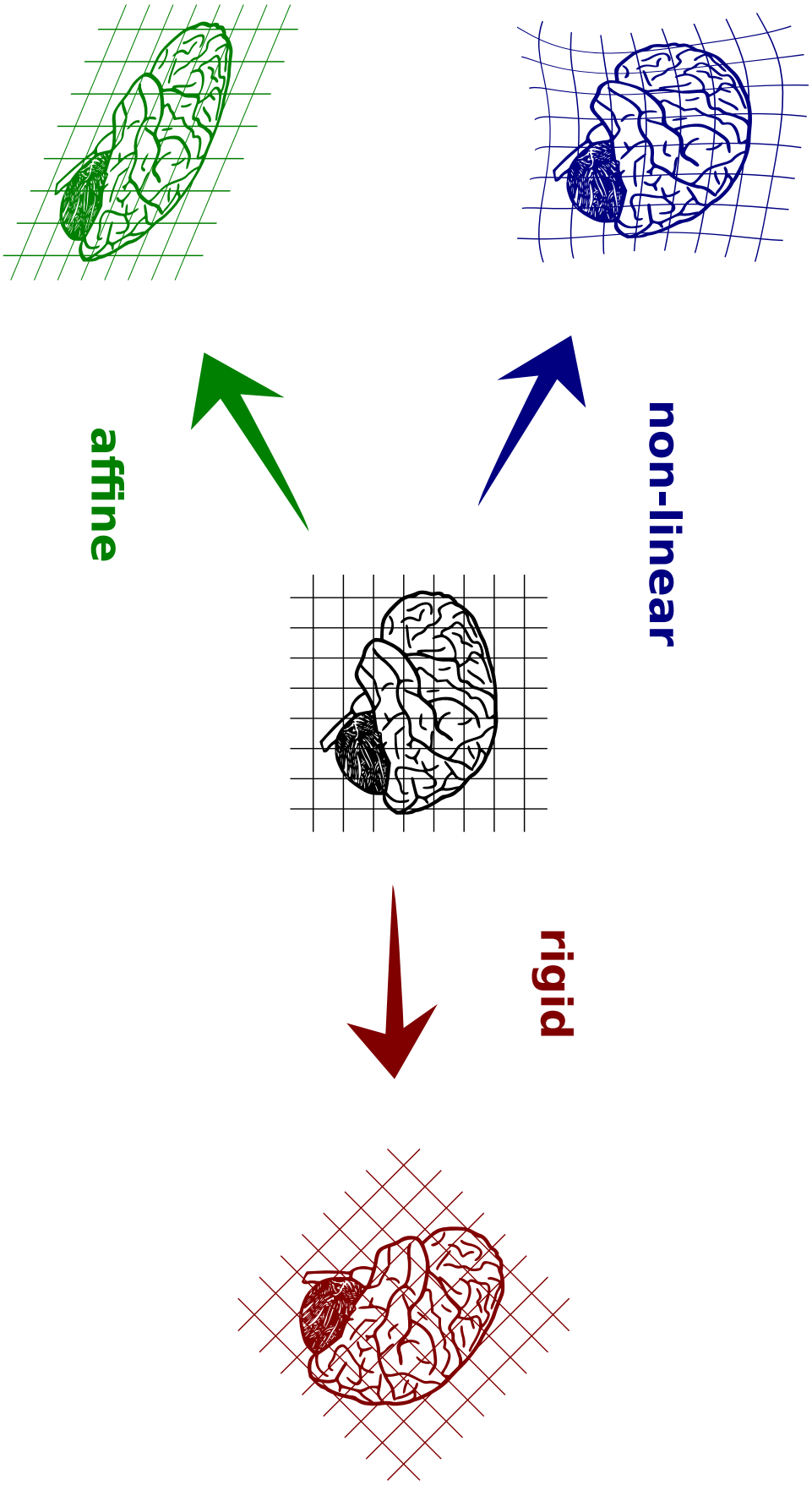

Figure 2.1: Parametrisations for image registration. Adapted with permission from the PhD thesis of Jonathan Young.

cortex. Different anatomical partitions, known as atlases, may be appropriate for the analysis of different anatomical changes [78]. The process of assigning each voxel in a brain image an anatomical label is commonly called parcellation, and belongs to the family of segmentation tasks. If one is using a pre-existing reference image constructed for groupwise comparisons, then this image is likely to already be manually labelled, and parcellation is not a problem. If, however, one wishes to take some regional measurements from an image without subjecting it to the degradation resampling, or one wishes to use a sample specific groupwise template of the type described in section 2.4.1, then one needs to find a way to provide regional labels for the voxels of brain images in their native state. Manual labelling is the gold standard method, but it requires expert knowledge and is too time consuming to be conducted on a large scale.

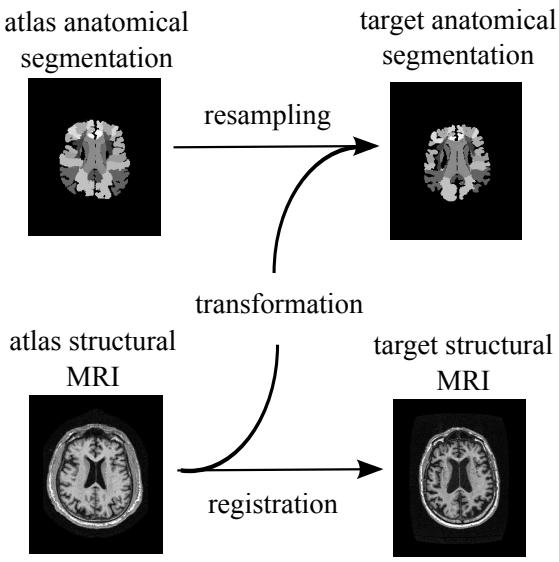

Figure 2.2: Atlas propagation using sMRI. Taken with permission from the PhD thesis of Jonathan Young.

To replace manual labelling, one may use **atlas propagation**. Atlas propagation methods automatically propagate anatomical labels from small sets of manually labelled images to the larger samples required in studies. In the simplest case, a single labelled image can be registered to a target and the resulting transformation used to resample the labels. This process is illustrated in figure 2.2. The success of this labelling will be dependent on a high degree of morphological correspondence between the labelled and target images. In order to improve the quality of the final parcellation, one may propagate labels from multiple labelled images, and then combine them with some form of label fusion, as illustrated in figure 2.3. This may be a simple majority vote, or a more advanced technique that takes the local correspondence between the labelled and target images into account [79]. Other techniques that can be employed to improve automatic parcellation include the pre-selection of the subset of labelled images that

are morphologically closer to the target, or even the use of images in the target sample that have already been parcellated as potential sources of labels for the remaining targets [80, 81].

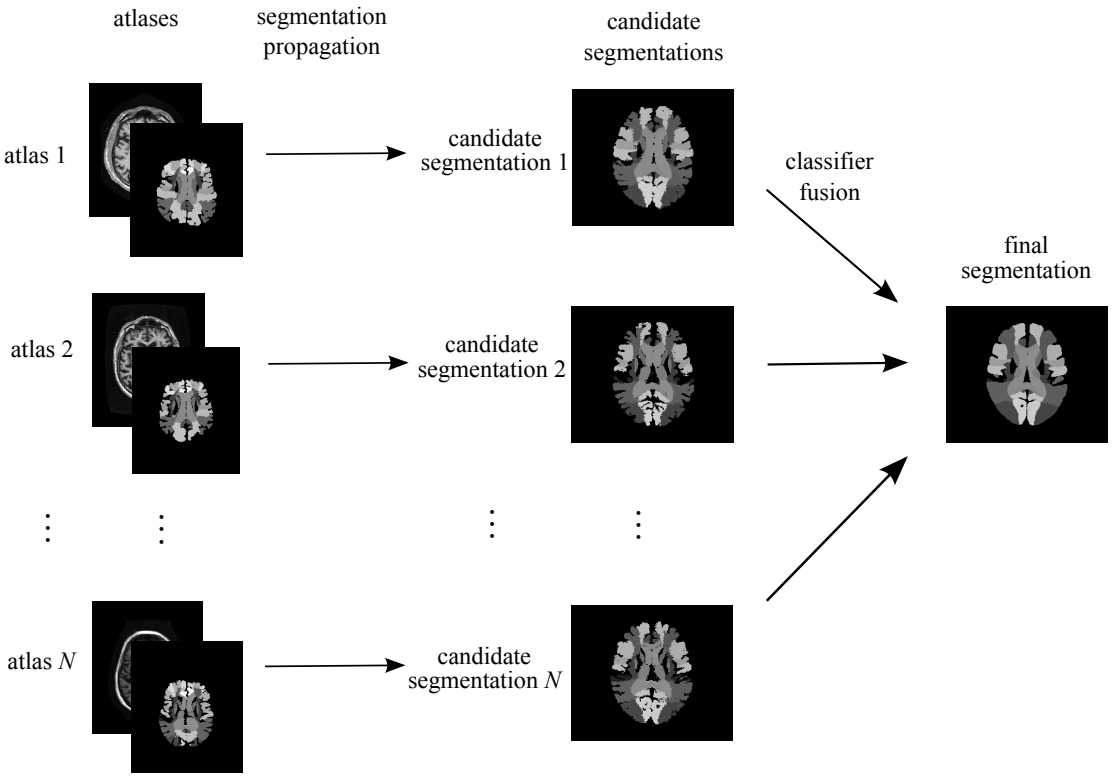

Figure 2.3: Atlas fusion. Taken with permission from the PhD thesis of Jonathan Young.

2.4.3 Tissue segmentation

.

While atlas parcellation divides the brain into anatomical regions, tissue segmentation divides it into physiological tissue classes, commonly grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Even more than atlas parcellation, tissue segmentation is reliant on a precise anatomical detail, and so can only be feasibly performed using sMRI. As with atlas parcellation, manual segmentation of the brain is too laborious to be practical in large scale studies, so an automatic method is necessary. The most common method for tissue segmentation uses a Gaussian mixture model [82], in which the intensities of the voxels from each tissue class have some Gaussian distribution with unknown parameters. Because a voxel may contain multiple tissue classes, the voxels are allowed fractional membership of each class, rather than being limited to one. The fraction of a tissue in a voxel is commonly referred to as a 'density' or 'concentration'. An expectation maximisation algorithm is commonly used to optimise the parameters of the distribution to maximise the *a posteriori* probability of the intensity distribution associated with the image. This works by alternatively updating the parameters of the

distributions and the memberships of the voxels until convergence is obtained.

Because the intensity distributions of the tissue classes have significant spatial overlap, it is necessary to introduce anatomical knowledge to regularise the problem in the form of a map detailing the *a priori* probabilities of each tissue class at every voxel. This map must be built from a set of manually labelled reference images, and propagated into the space of the target image using registration. This works well, except in the case where the target image is too morphologically dissimilar from those used to build the prior [83]. Additionally, the intensities of the voxels of a given class, while locally similar, may be modulated across the image by differences in magnetic field strength. To overcome this, parameters describing a 'bias field' (of differences in magnetic field strength) are incorporated into the model and simultaneously optimised. Finally, in order to exploit the fact that the tissue classes typically lie in contiguous regions, a penalty term is introduced into the model to make it more likely for voxels' neighbour to have the same tissue class. Because this penalty only applies to local interactions, it effectively models tissue concentrations as a discrete Markov random field. An illustration of the tissue segmentation is provided in figure 2.4.

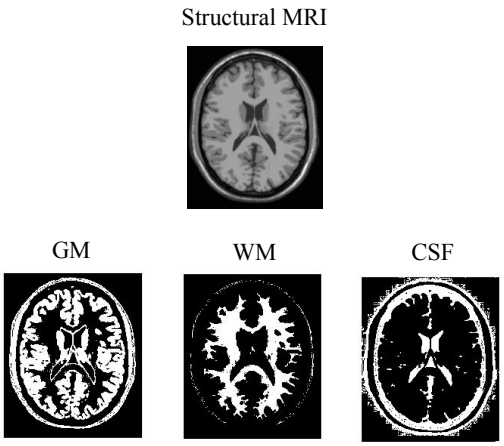

Figure 2.4: A grey matter, white matter and CSF tissue segmentation. Taken with permission from the PhD thesis of Jonathan Young.

The comparison of tissue concentrations (typically GM) on a voxelwise basis may be called voxel-based morphometry (VBM) [84]. The VBM techniques that have become established for group differences studies are also commonly used in ANA research. After tissue concentrations have been produced, registration is used to propagate them to the groupwise template for comparison. Because differences in region volume may be removed by registration, it is necessary to adjust the tissue concentrations accordingly. This is done by modulation using the Jacobian determinant map, which describes the local degree of the contraction/expansion

introduced by registration. In this way, the total amount of tissue concentration in a region is conserved. Finally, to accommodate small differences in the precise location of anatomy, and to account for the spatial variability in the effect of disease, a spatial smoothing (i.e., Gaussian convolution) is commonly applied.

2.4.4 Cortical thickness measurement

Cortical thickness measurement [85, 86] appears very often in ANA for AD [24, 29]. Using an automatically generated tissue segmentation, one of two methods is used to estimate the thickness of GM comprising the cortex at many vertices across the exterior surface of the cerebrum. This map of thickness may provide greater sensitivity to changes than the use of the GM concentrations alone [87]. The first method is based on the fitting of mesh surfaces. Briefly, using an appropriate degree of regularisation, one fits one mesh surface to the GM/CSF outside of the cortex, and another to the GM/WM on the inside [85]. The distance between the two can then provide a measure of thickness. In the second, voxel-based, method one uses a physical model in which the WM inside is held at a certain electrostatic potential, and the CSF outside is held at another. The field lines of the resultant electric field are then found by solving Laplace's equation [86]. They form a natural trajectory between the outer and inner surfaces of the cortex, as they take a short path between the outer and inner surfaces that produces a one-to-one mapping between the locations on both. The length of these lines is then taken as the cortical thickness. As with voxel-based studies, some degree of smoothing may be applied to ensure anatomical correspondence in comparative studies.

2.5 Imaging features and feature reduction

In ANA, the purpose of the tools described in the previous section is to produce a set of meaningful features capable of describing disease. In some cases, these features may be considered to contain a mixture of informative signal and some unwanted 'noise' variability that may confound the construction of effective models. In this case, one may wish to discard the noise component of the original features to produce a new lower dimensional representation using **dimensionality reduction**.

This section describes the various choices of features that may be used, and some of the most common dimensionality reduction methods.

2.5.1 Features

For sMRI, one of the oldest and most common choices of feature set is the voxel GM (or, less commonly, WM [88]) tissue concentrations borrowed from VBM [24, 34, 89]. Either the whole brain or only a subregion (such as the cerebrum, temporal lobe) may be used [88]. If

spatial normalisation is considered undesirable, tissue concentrations can be compared using regional measures instead [46]. To do this, one must perform an atlas propagation to bring regional labels to all images in the sample, and find the sum of the tissue concentrations in each region [90]. In doing this, it is important to control for the different resolutions of the images in the sample, and the different head sizes of the subjects as measured using total intra-cranial volume (TIV). An even simpler feature set may be produced using the volumes of the regions themselves [90]. Again, in this case, it is important to account for difference in TIV. Because of its role in AD, the hippocampus receives particular attention as a discriminating feature. Both hippocampal volume and shape, as quantified using spherical harmonics, have been used in multiple ANA studies [88, 91, 92]. Lastly, another common measure from sMRI is the cortical thickness measurements described in section 2.4.4 [44, 49]. Thickness may be compared on a per-vertex basis, or by looking at the thickness in anatomical regions [88]. The number of vertices is typically of the order of tens or hundreds of thousands.

For nuclear imaging modalities such as FDG-PET, a common choice of features is the use of voxel intensities [28, 93]. This is similar in many respects to the use of tissue concentrations from sMRI, but involves an additional problem of normalisation. Because the level of the voxel intensities will depend heavily on a variety of factors that are hard to control for, it is more informative to look at their relative, rather than absolute values. To do this, one must define a reference region whose mean is used to rescale the intensities. This region is ideally one whose physiology should be relatively constant across the sample, requiring that it be unaffected by disease. For FDG-PET, a common choice of reference region is cerebellar GM, though there are other feasible alternatives [94]. Just like tissues concentrations, nuclear image voxel intensities may also be compared on a regional basis [57] where an anatomical parcellation is available.

Resting state fMRI is used to provide measures of functional connectivity between anatomical regions, which can be used as features in ANA [95–97]. These are typically measured as the temporal correlations in blood-oxygen-level dependent contrast after applying a frequency filter. Alternatively, diffusion tensor imaging can be used to provide features describing structural connectivity [95]. In this case, the connection between two regions is derived from the amount of the WM tracts that link them. Local summary measures such as the mean diffusivity or the fractional anisotropy can also be used to provide a local description of diffusion [98].

When voxel-based features (tissue concentrations, nuclear imaging intensities) are used, the number of features used will commonly be very high (tens or hundreds of thousands), particularly when using high resolution images. The number of regions in an anatomical parcellation is typically of the order of one hundred [78, 99, 100], which leads to a number of pairwise con-

nections of the order of ten thousand. Typical samples are of the order of one hundred subjects, with many studies having smaller samples. This means that in the majority of cases, the number of features is greater than the number of training examples. This can make it very difficult to fit complicated (e.g., non-linear) prediction models with good generalisation beyond the training set. This is another motivation for dimensionality reduction.

2.5.2 Dimensionality reduction

Dimensionality reduction methods may be divided into two families: feature extraction and feature selection. Feature extraction approaches produce some transform from the full set of features to another set with lower dimensionality. A given feature of the new set may be related to multiple features in the original one. Feature selection approaches select some subset of the features that is intended to contain only those that are most relevant.

2.5.2.1 Feature extraction

Perhaps the most simple feature extraction method seen in ANA is principal component analysis, which entails the projection of the features into a linear subspace of fixed dimensionality which best preserves their covariance structure [55, 93]. A related method is partial least squares [55, 101], this method seeks to preserve the covariance between the features and the labels to be predicted, rather than the covariance of the features with themselves. In addition to these linear methods, there are a variety of non-linear methods including Laplacian eigenmaps [102], locally linear embedding [103] and stacked auto-encoders [104].

2.5.2.2 Feature selection

One simple approach to feature selection is to use mass univariate testing to identify those features with the greatest apparent association with the labels [105, 106]. Alternatively, one may use sparse regression techniques [103] based on l_1 regularisation. A more direct approach is recursive feature elimination, which uses CV to see which features may be eliminated without a decrease in performance [105]. One may also consider a knowledge-driven selection of features. In AD ANA, one may choose to use only those imaging features derived from the temporal lobes, as these are regions known to be strongly affected by the disease. There is a sense in which supervised learning algorithms themselves perform data-driven feature selection internally; an effective algorithm will be able to identify which features should be used when there is a sufficiently large number of examples. It may be for this reason that data-driven methods did not appear to convey much advantage in a 2012 study on AD classification by Chu et al. [105]. In that study, only knowledge-driven feature selection provided a significant advantage, and this declined with increasing sample size.

2.6 Machine learning algorithms

This section describes learning algorithms that may be used in ANA. Again, I have chosen the methods described based on their importance to the field and their appearance in my work. I have divided them into several groups based on computational considerations.

2.6.1 Kernel methods

As discussed in the previous section, many common imaging feature sets have very large numbers of dimensions. This can make supervised learning very demanding in terms of computation and memory. Fortunately, many learning algorithms do not need to work with the **primal** or original representation of the data. Instead, they may work with a **dual** representation that is much lower in dimension. This dual representation, derived from the primal, may take the form of a **kernel** or Gram matrix of inter-point dot products, or of the matrix of squared pairwise inter-point distances. Either one of these representations can be derived from the other by a simple transformation. Where the sample of n items is represented by n vectors of length d, the primal representation of the sample will contain nd values, while the dual representation requires n^2. As n is typically much less than d, this can provide a great gain in performance. This is particularly important in the work presented later in this thesis.

Because they represent dot products and square distances, the kernel and distance matrices of several feature sets may be combined additively. In order to adjust the relative contributions of each original feature set to the variability of the data, the kernels should be multiplied by some relative weights before they are added. Various procedures exist to determine the relative weights automatically. The use of these methods to combine various feature sets is known as multi-kernel learning. In medical imaging, the different kernels may be derived from different imaging modalities [28, 46, 52], or different anatomical regions [107].

Another important aspect of kernel methods is that they allow for something called the kernel trick, where linear methods can be extended to make non-linear ones. Rather than using the standard definitions of the dot product or inter-point distance, one may replace these with some surrogate $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ called a kernel function. In the context of kernel methods, the standard dot product may be called the linear kernel. Where a learning algorithm that produces linear decision functions can be specified in terms of the standard kernel matrix, the use of a kernel matrix produced by the surrogate function can extend the algorithm to produce non-linear decision functions. The most common surrogate function is the radial basis function (RBF) defined

$$K(X_1, X_2) = \exp\left(-\frac{||X_1 - X_2||^2}{2\sigma^2}\right),$$

where X_1 and X_2 are the vector valued features of two items, and σ is a free parameter. This will be illustrated in the description of the support vector machine.

2.6.1.1 The support vector machine

The support vector machine (SVM) is a family of learning methods derived from statistical learning theory, specifically empirical risk minimisation [108, 109]. It is now by far the most popular learning algorithm in ANA research [24, 29, 32]. In binary classification tasks where $\mathbb{Y} = \{-1, +1\}$, the SVM produces a decision function parametrised by a 'fat plane' comprising two parallel bounding hyperplanes and some separating width. This fat plane is parametrised in the primal case by a weight vector w and an offset b. The bounding hyperplanes are the sets of points $x \in \mathbb{X}$ satisfying $x \cdot w - b = -1$ and $x \cdot w - b = +1$ respectively. Unlabelled items with features x are assigned the label $\text{sign}(x \cdot w - b)$.

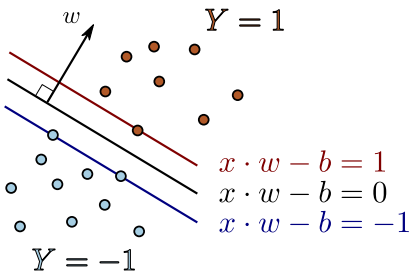

Figure 2.5: The fat hyperplane of an SVM.

The hyperplane is selected as a compromise between maximising width (achieved by minimising $||w||$) and successfully separating of the points of the training set associated with the different labels as illustrated in figure 2.5. This is justified by a result from empirical risk minimisation that states that the difference between the fat plane's success in classifying the items of the training set and its success in classifying unseen items is limited by its width. Computationally, the hyperplane is determined by minimising a convex cost function. This includes a regularisation term C that controls the trade-off between width and successful separation of the training items. This cost function can be reformulated so that the items of the training set appear only in the kernel function, allowing surrogate kernel functions such as the RBF to be used. The impact of the kernel function in SVM classification is illustrated in figure 2.6.

Practical concerns. The SVM has been generalised to a variety of tasks including density estimation and regression, but it still best known as a classification tool. The SVM does not natively support multi-class classification, but various methods exist to extend binary classifiers to this case [110]. As it does not entail a generative model, it cannot provide a truly probabilistic output, though various extensions exist to provide this function [111]. Another practical concern

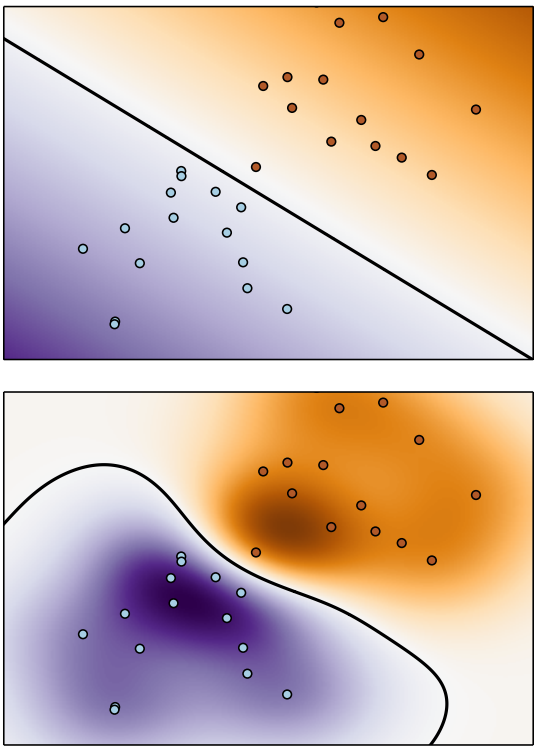

Figure 2.6: Linear (above) and RBF kernel (below) SVM classification. The black line represents the boundary where $x \cdot w - b = 0$. Unseen items will be classified based on which side of it they fall.

in SVM classification is the selection of C. A common approach is to use CV to estimate the performance of the SVM on the sample at a variety of possible C values, and then select the one that seems the best. One occasionally sees statements to the effect of 'we use the default C value of 1.0'. However, due to the interaction between C and the scale of the problem (magnitude of the kernel values), there can be no meaningful default value, as C values are not comparable between problems.

When the RBF kernel function is used, its σ parameter poses a similar problem. In the very high dimensional regime where the number of dimensions is greater than the number of available training items, any two groups of items will be linearly separable. In this case, perfect training set classification (achieved with very high C) will always possible. It is not intuitively obvious that RBF kernels should convey any benefit when the data are already linearly separable, as they introduce further degrees of freedom into a problem that is already under-constrained.

2.6.1.2 Relevance vector machines and Gaussian processes

The relevance vector machine (RVM) was developed to avoid the issue of parameter selection in SVM classification and to produce a more naturally probabilistic output [112]. The RVM is based on an extension of Bayesian regression that enforces sparsity in the dual space represen-

tation of the weight vector. While RVMs avoid the need to select free parameters using CV, the optimisation they entail is less simple than that needed for SVMs. An ANA specific extension of the RVM has been developed to exploit voxel-based features [89] while promoting both sparsity and spatial smoothness in the primal weight vector. Gaussian process classification and regression is another probabilistic kernel method of note [28, 113], of which the RVM can be seen as a particular specialisation.

2.6.1.3 Linear discriminant analysis

Though it might not traditionally be considered a kernel method, linear discriminant analysis (LDA) can be expressed in a dual form. Briefly, LDA is based on the projection of the training data that maximises the ratio of the inter-class variance to the intra-class variance [114]. For the sake of simplicity and stability in the estimation of the covariance, the distributions associated with each class are assumed to have the same covariance structure. This projection produced by LDA may be used as a method of feature reduction, or directly for classification [115]. While LDA is commonly seen as being a relatively simple and ancient method (one version was invented by Fisher in 1936), a large amount of recent research has been conducted on the appropriate degree of regularisation to use in the estimation of the covariance [116]. LDA is perhaps the most common learning algorithm in AD ANA after the SVM [29].

If the assumption that the class distributions have identical covariance structures is relaxed, then the decision functions produced are no longer linear in the primal space, and the resulting procedure is instead called quadratic discriminant analysis (QDA). QDA may provide better results than LDA where the covariance structures can be estimated with high accuracy [117]. QDA has been used in various ANA studies, though it is much less common than LDA [29].

2.6.2 Simple or toy algorithms

This section will describe several simple learning algorithms that are not commonly used in ANA but which do feature in the work of this thesis. I have chosen them for their computational efficiency, and I use them to study the behaviour of validation strategies for general learners and algorithms.

2.6.2.1 K-nearest neighbours

K-nearest neighbour (KNN) methods use a representation of the distances between items. In classification tasks, new items are classified by a majority vote of their K nearest labelled neighbours. In regression tasks, labels are predicted by a weighted sum over the labels instead. KNN methods have some strong consistency results [118] and can be efficiently cross validated [119]. KNN is non-linear and can be straightforwardly applied to learning problems in which features

are specified by a kernel matrix.

2.6.2.2 Naive Bayes

Naive Bayes (NB) classification entails a simple model in which all features are assumed to be independent [120]. For continuous features, each class is assumed to have a Gaussian distribution whose parameters are estimated using maximum likelihood. For categorical features, each class is assumed to have a discrete distribution where the probabilities associated with each feature value are estimated similarly. After the distributions of each class have been estimated, one can simply apply Bayes' rule to estimate the posterior probability of an unlabelled item belonging to each class. Where a categorical label estimate is required, rather than a probabilistic one, NB simply outputs the most likely class.

2.6.2.3 Nearest Centroid

The nearest centroid (NC) classification algorithm uses the training set to estimate the centroids (expected value of X) associated with the distributions of each class of items. Unlabelled items are assigned the class of the centroid that is nearest to them. NC can be applied with kernel features.

2.6.2.4 C45 decision trees

The C45 algorithm was developed for data mining and classification by Quinlan in the early 90s [121]. The algorithm builds a decision tree by recursively dividing the feature space into disjoint subsets. The divisions are selected so as to minimise the entropy associated with the distribution of the labels in each subset. A heuristic based on confidence intervals for the predictive accuracy of the tree is used to halt the division before the number of items remaining in each subset becomes too small. This limits the depth of the tree.

2.6.3 Ensemble methods

Ensemble methods work by combining the results of many "weak" predictors constructed using some type of random perturbations of a learning algorithm or training set. In the case of the bagging ensemble technique, this entails building each weak predictor with a bootstrap-resampled version of the training set [122]. The weak predictors may individually have worse performance than they would in the absence of any perturbation, but when their predictions are combined, they can provide a performance greater than was possible for a single predictor in the unperturbed case. By averaging over many perturbed predictors, it may be possible to smooth out the unstable (and hence unreliable) aspects of a predictor's decision function. In this way, ensemble methods allow the use of flexible learning algorithms that can provide a rich space of decision functions without the high risk of over-fitting.

2.6.3.1 Random forests

Random decision forests are a turn of the century ensemble method for classification and regression. They are based on bagged, randomised decision trees [123]. Since their invention, they have been extended and generalised to a variety of tasks [124]. Though they are not common in AD ANA research [24], they are perhaps one of the more common non-linear methods [22, 45, 60, 125]. In my work, they appear as an example of a state-of-the-art alternative to the more common linear methods. An illustration of the combination of random decision trees is presented in figure 2.7.

2.6.3.2 Other ensemble methods

Other ensemble methods seen in AD ANA include SVM ensembles [126] and boosting [127].

2.6.4 Others

Other learning algorithms of note include logistic regression [128, 129] and artificial neural networks such as the extreme learning machine [130] or multi-layer perceptron [131]. More recently deep neural networks, which involve many layers between the input and output variables, are being used [104, 132, 133].

Summary

In this chapter, I have reviewed the key problems, materials and methods of ANA research. I have discussed the types and origins of the imaging data ultimately used to produce features, and the tools with which this is done. I have introduced some of the most important learning algorithms that are used to build models linking these features to clinical variables, including all that are used in this thesis. The next chapter will discuss the cross validation through which ANA methods must be ultimately validated.

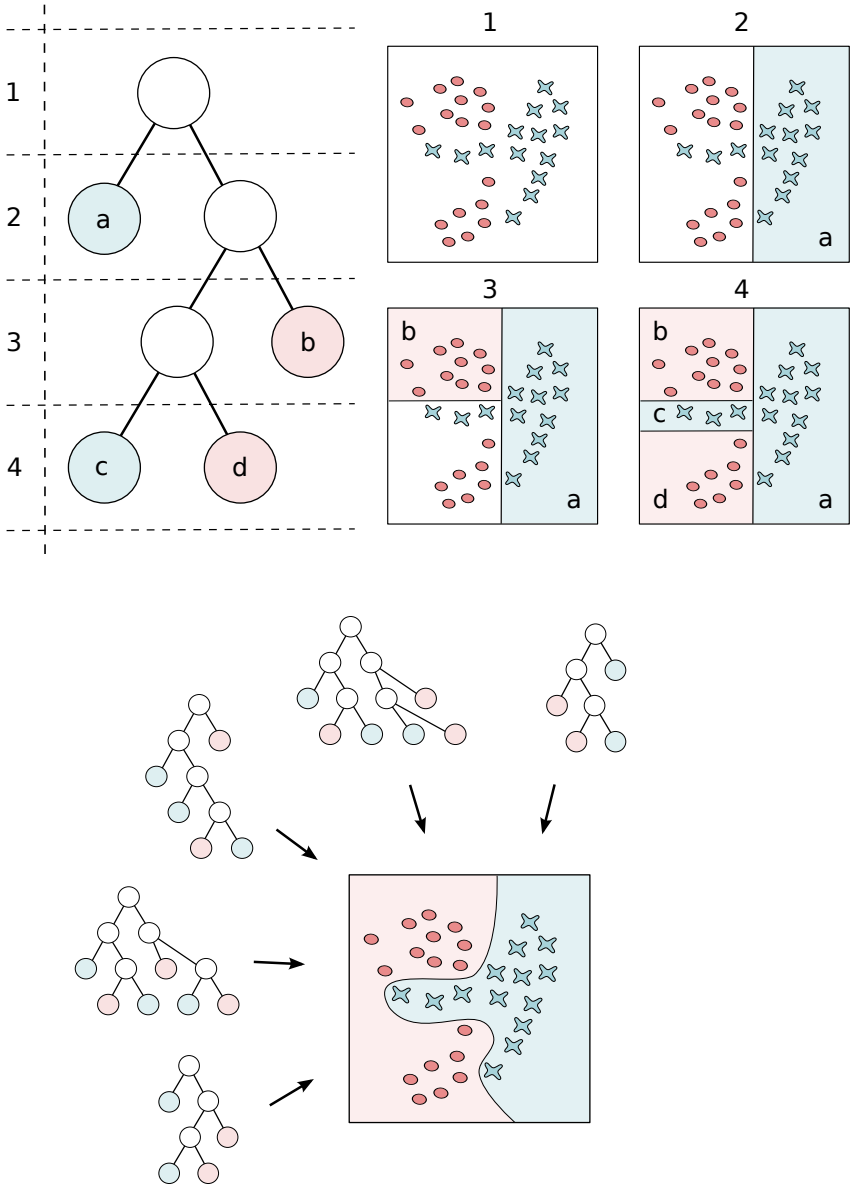

Figure 2.7: Above, a single decision tree whose nodes represent different partitions of the feature space creates a decision function to classify future items. Below, the decisions of multiple perturbed decision trees are combined to implement a random forest.

Chapter 3

Key concepts in performance measurement

In this chapter, I shall introduce the key concepts required to understand supervised learning and cross validation (CV) in ANA. I shall begin with a mathematical description of supervised learning and the ideal experiments that define performance quantities. I shall then describe the resubstitution and CV experiments that must be used to estimate performance in practice, and why the latter is more suitable in ANA. I shall provide an outline of the common CV strategies, and discuss stratification with item subpopulations. I shall introduce two important issues that will be addressed in this thesis: the problem of dependency between component CV results and the problem of selection bias and over-fitting in model selection. Finally, I shall review the purpose and practice of CV in AD ANA.

In this chapter and all those that follow, let

$$\mathbf{a} = \langle a_i \rangle_{1 \leq i \leq n} \tag{3.1}$$

denote a sequence of length $n \in \mathbb{N}$ whose ith value (where $1 \leq i \leq n$) is a_i. Where each a_i is a member of the set $\mathbb{A}$, the sequence $\mathbf{a}$ is a member of the set $\mathbb{A}^n$. The set of non-empty arbitrary length sequences of values in $\mathbb{A}$ is denoted

$$\mathbb{A}^+ = \bigcup_{i \in \mathbb{N}} \mathbb{A}^i . \tag{3.2}$$

3.1 Basic concepts

I shall call the atomic observations that comprise the data in prediction problems **items**. Each item $W = (X, Y)$ is an ordered pair of two variables: some descriptive **features** that are always available (denoted $X \in \mathbb{X}$), and some dependent **label** (denoted $Y \in \mathbb{Y}$). Items may be **unlabelled**, meaning that while the label exists, it is hidden. Typically, W is a random variable with some distribution in the joint feature-label space $\mathbb{W} = \mathbb{X} \times \mathbb{Y}$. It is the goal of supervised

learning to explicitly or implicitly model this distribution so that the labels of unlabelled items can be predicted. This is done using a prediction function or **predictor** $t : \mathbb{X} \to \mathbb{Y}$. The set of predictors is simply the set of functions $\mathbb{X} \to \mathbb{Y}$, denoted $\mathbb{T}$.

A predictor may be pre-constructed and derived from prior knowledge, in which case it may be regarded as a fixed, immutable object. Alternatively, it may have to be inferred from a sequence of labelled items called the **training set**, denoted $\mathbf{G} \in \mathbb{W}^+$. Despite its name, it is not a set in the rigorous sense, but a sequence of random variables which may itself be regarded as a random variable. In order to produce a predictor, one must use a **learner** $u : \mathbb{W}^+ \to \mathbb{T}$. The set of learners is the set of functions $\mathbb{W}^+ \to \mathbb{T}$, denoted $\mathbb{U}$.

3.1.1 Realisation in ANA

In ANA, each item corresponds to a person. An item's features are some numeric description of the person's neuroimaging data, and its label is a description of the level or type of disease in the person. Where the labels are categorical (e.g., $\mathbb{Y} = \{\text{red}, \text{blue}, \text{green}\}$), they tend to describe the type or presence absence of disease, and the prediction problem is called **classification**. Where the labels are real valued (i.e., $\mathbb{Y} \subset \mathbb{R}$), they tend to describe the continuously varying severity of disease. In this case, the prediction task is called **regression**.

In many applications, the specification of the features, and thus also that of learners and predictors, may be ambiguous. Consider the common scenario in which a base set of imaging descriptors $X \in \mathbb{X}$ undergoes some processing step defined by a pre-specified function $f : \mathbb{X} \to \mathbb{X}'$ to produce the derived descriptors $X' \in \mathbb{X}'$ on which a standard learning algorithm (e.g., KNN or SVM) is applied. This processing step could be image processing, feature selection, dimensionality reduction, or anything else that does not require information related to the labels. (One could even consider image acquisition to be an unsupervised processing step. In this interpretation, X would represent the unknown biological state of a patient, and X' would represent an image.) There are two possible formalisms:

1. The feature space is $\mathbb{X}'$. A predictor is a function $t' : \mathbb{X}' \to \mathbb{Y}'$ in the space $\mathbb{T}'$, and a learner is a function $u : \mathbb{W}'^+ \to \mathbb{T}'$, where $\mathbb{W}' = \mathbb{X}' \times \mathbb{Y}$. In this formalism, learners are uniquely specified by the choice of standard learning algorithm.

2. The feature space is $\mathbb{X}$. A predictor is a function $t : \mathbb{X} \to \mathbb{Y}$ in the space $\mathbb{T}$, and a learner is a function $u : \mathbb{W}^+ \to \mathbb{T}$, where $\mathbb{W} = \mathbb{X} \times \mathbb{Y}$. In this formalism, the processing step a learner uses is incorporated into its specification. This formalism is potentially more complicated, but it is able to describe learners that make use of the base features without first applying the processing step f.

When comparing learners, one needs a formalism that has a feature specification that is broad enough to accommodate all of them. In ANA, the choices of scanner, image processing, and feature extraction method may all differ between pipelines. Accordingly, all these steps may have to be incorporated into the definition of a learner/predictor in a rigorous analysis. This scheme is illustrated in figure 3.1.

Where a processing step uses a function inferred from the data (as in common feature selection/reduction techniques), it must be considered part of the learner. The descriptors $\mathbb{X}'$ may no longer be considered as a feature space for the problem, as the distributions of items in it will differ depending on whether they were used in the construction of the transformation. For more detail on this issue, see appendix A.

3.2 Measuring performance

The **performance** or quality of a prediction $t(X)$ must be assessed by a utility metric $\phi : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$. For a random item $W = (X,Y)$, the performance of a predictor on that item is the random variable $\phi\big(Y,t(X)\big)$. In this work, I take the convention where higher values of performance are desired. One could also use a convention where the utility of a prediction is measured in terms of an *error* rather than a *performance*. Under that convention, low error values would be considered desirable.

In classification, where the labels are categorical, the typical ϕ is the **accuracy** metric defined

$$\phi(Y,\hat{Y}) = \begin{cases} 1 & \text{if } Y = \hat{Y} \\ 0 & \text{otherwise.} \end{cases}$$

In regression, where the labels are real valued, one may use the negative squared error metric, defined $\phi(Y,\hat{Y}) = -(Y-\hat{Y})^2$.

In order to associate a performance with learners and predictors, it is necessary to define two abstract experiments involving randomly generated sequences of items.

3.2.1 The testing experiment for a fixed predictor

In a testing experiment, a fixed predictor t is evaluated on a **testing set** $\mathbf{H} := \langle H_i \rangle_{1 \le i \le n}$ in the space $\mathbb{W}^+$. This is a sequence of n random items (n may be fixed or a random variable) which may itself be regarded as a random variable. The performance measured on the ith item of the testing set is the random variable

$$Q_i = \phi\Big(Y_i, t(X_i)\Big), \text{ where } H_i = (X_i, Y_i)\,, \tag{3.3}$$

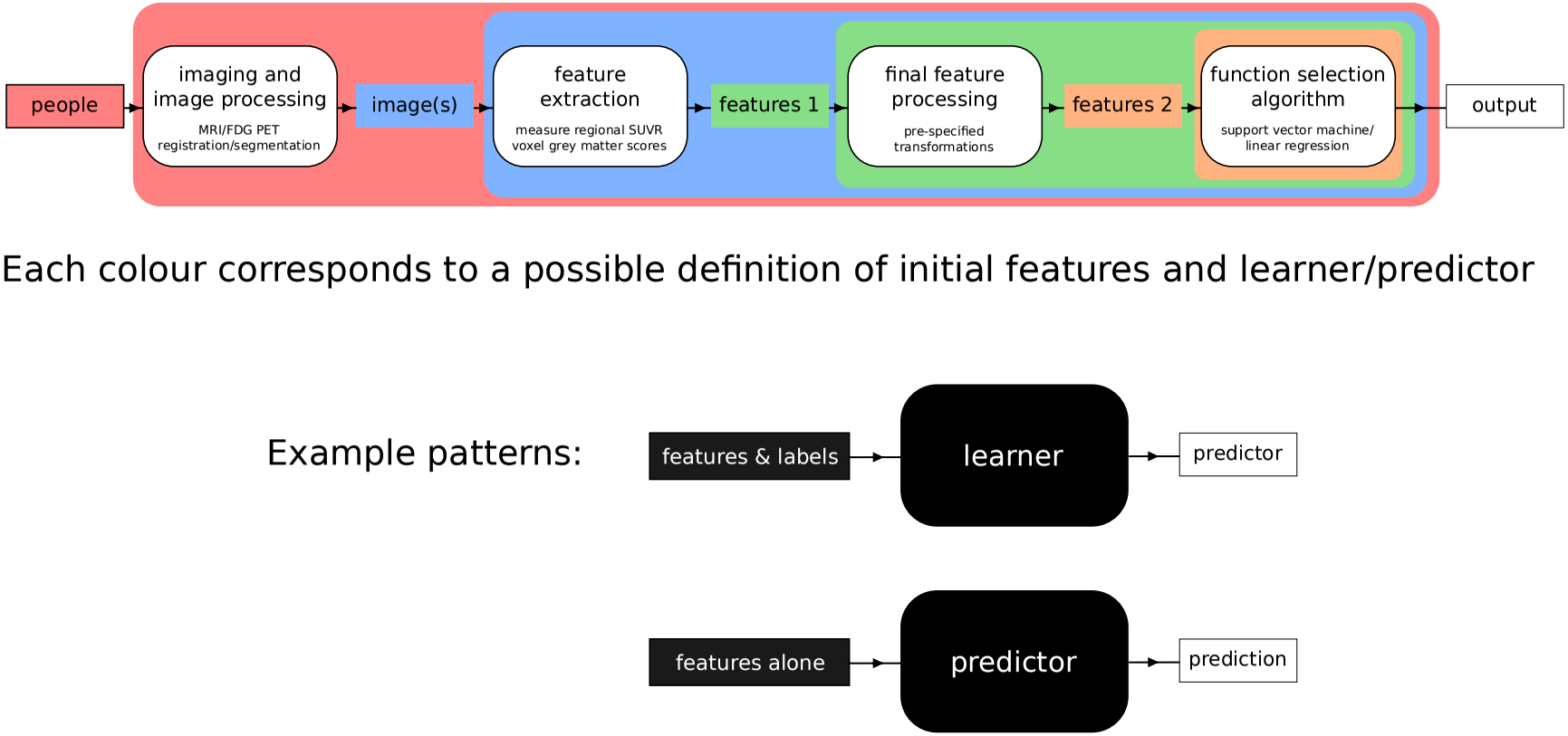

Figure 3.1: Ambiguity in the learner/predictor definition resulting from the specification of the features (as described in section 3.1.1). Each small coloured box represents a possible definition of the features in a learning problem; the larger coloured box of the corresponding colour surrounds the set of actions that implement the learner/predictor under that feature definition.

and the mean performance on the whole testing set is the random variable $\bar{Q}$, defined

$$\bar{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i \,. \tag{3.4}$$

One can define the predictor evaluation function $g : \mathbb{T} \times \mathbb{W}^+ \to \mathbb{R}$, which takes the predictor t and the testing set $\mathbf{H}$ as its inputs and returns $\bar{Q}$ as defined above as its output.

The **predictor performance** associated with t is simply the expected performance in the testing experiment, denoted $\mu_t = \mathbb{E}_{\mathbf{H}}[\, g(t,\mathbf{H})\,]$. This will depend on the precise distribution of the testing set $\mathbf{H}$. When the items of $\mathbf{H}$ are independently drawn from a single generating population, then their distribution and number fully specifies the distribution of $\mathbf{H}$. In this case, μ_t is independent of n, as the expected performance on the testing set is the same as that expected on any one of its items.

3.2.2 The train-test experiment

The train-test experiment involves both random training and testing sets, denoted $\mathbf{G}$ and $\mathbf{H}$ respectively. Rather than taking the fixed value t, the predictor produced by a learner u in a train-test experiment is the random variable $T = u(\mathbf{G})$. The performance of this predictor is assessed on the testing set in exactly the same way as in the testing experiment for a fixed predictor, to produce the random variable performance measurement $g(T,\mathbf{H})$. One can define the learner evaluation function $\gamma : \mathbb{U} \times \mathbb{W}^+ \times \mathbb{W}^+ \to \mathbb{R}$ which takes a learner $u \in \mathbb{U}$, uses it to select a predictor $T = u(\mathbf{G})$ on a training set $\mathbf{G}$, and evaluates it on a testing set $\mathbf{H}$. By definition,

$$\gamma(u,\mathbf{G},\mathbf{H}) := g\big(u(\mathbf{G}),\mathbf{H}\big) := g(T,\mathbf{H})\,. \tag{3.5}$$

The **learner performance** associated with a learner u is the expected performance in the train-test experiment, defined

$$\mu_u = \mathbb{E}_{\mathbf{G},\mathbf{H}}[\, \gamma(u,\mathbf{G},\mathbf{H})\,]\,. \tag{3.6}$$

Just like the predictor performance, the learner performance is unaffected by n when the items of $\mathbf{H}$ are i.i.d. from a single generating population.

The expected result in such an experiment, conditional on the training set $\mathbf{G}$, is the random variable

$$\begin{aligned} M_T &= \mathbb{E}_{\mathbf{H}}\big[\, \gamma(u,\mathbf{G},\mathbf{H}) \,|\, \mathbf{G}\,\big] \\ &= \mathbb{E}_{\mathbf{H}}\big[\, g\big(u(\mathbf{G}),\mathbf{H}\big) \,|\, \mathbf{G}\,\big] \\ &= \mathbb{E}_{\mathbf{H}}\big[\, g(T,\mathbf{H}) \,|\, \mathbf{G}\,\big]\,, \end{aligned} \tag{3.7}$$

which is the predictor performance of T in the case where $\mathbf{G}$ and $\mathbf{H}$ are independent. For a predictor u, the distribution of $\mathbf{G}$ specifies the distribution of T, and hence that of M_T. By definition, the expectation of M_T is the learner performance μ_u:

$$\begin{aligned}\mathbb{E}_{\mathbf{G}}[M_T] &= \mathbb{E}_{\mathbf{G}}\Big[\mathbb{E}_{\mathbf{H}}\big[\gamma(u,\mathbf{G},\mathbf{H})\,|\,\mathbf{G}\big]\Big] \\ &= \mathbb{E}_{\mathbf{G},\mathbf{H}}\big[\gamma(u,\mathbf{G},\mathbf{H})\big] \\ &= \mu_u\,.\end{aligned} \tag{3.8}$$

When $\mathbf{G}$ contains a fixed number m of i.i.d. items, then its distribution is fully determined by m. This is enough to fully specify the distribution of M_T, which determines the expected performance estimate in a train-test experiment where $\mathbf{H}$ contains an arbitrary $n \geq 1$ i.i.d. items. As a consequence, when both $\mathbf{G}$ and $\mathbf{H}$ are composed of a fixed number of i.i.d. items, m is sufficient to determine the expected performance result of the experiment (i.e., the learner performance μ_u). This means that, for a specified context with i.i.d. items, the learner performance μ_u may be regarded as a function of m.

Typically, μ_u will increase with m, as the greater information provided by larger samples should facilitate the selection of more effective predictors. The rate of increase will generally diminish with m, due to the diminishing marginal information provided by additional items.

3.3 Practical performance measurement

In real experiments, the data consists of a single sequence $\mathbf{D} \in \mathbb{W}^+$ of l random items. These items are commonly regarded as being i.i.d. from a single generating population.

There are two quantities one may wish to measure:

The full sample predictor performance. This is $\mathbb{E}_{\mathbf{H}'}[g(u(\mathbf{D}),\mathbf{H}')]$, where $\mathbf{H}'$ is some independent set of $n' \geq 1$ i.i.d. items from the same population as those in $\mathbf{D}$. This reflects the expected utility of the predictor constructed using the maximum possible training sequence size in a future application.

A representative learner performance. This is $\mathbb{E}_{\mathbf{G}',\mathbf{H}'}[g(u(\mathbf{G}'),\mathbf{H}')] = \mathbb{E}_{\mathbf{G}',\mathbf{H}'}[\gamma(u,\mathbf{G}',\mathbf{H}')]$, where $\mathbf{H}'$ is defined as before, and $\mathbf{G}'$ an independent set of some m' i.i.d. items. This quantity tells one something about the utility of a learner in a context independent of any particular realisation of a training set.

There are, broadly speaking, two ways to estimate these quantities: **resubstitution** and **cross validation**.

3.3.1 Resubstitution

Resubstitution is the use of the full dataset for both training and testing. The resubstitution performance estimate may be expressed $\gamma(u, \mathbf{D}, \mathbf{D}) = g(u(\mathbf{D}), \mathbf{D})$. Though the items of $\mathbf{D}$ are independently and identically generated from the same population as those of the testing set $\mathbf{H}'$ that defines the full sample predictor performance, they are no longer representative of that population for the purpose of evaluating the performance of the predictor $T = u(\mathbf{D})$. Due to their use in its construction, the predictor $u(\mathbf{D})$ has a special relationship with the items of $\mathbf{D}$; it is peculiarly well suited to predicting their labels, as it was selected by the learner for precisely that property. This gives the resubstitution performance estimator an optimistic bias when used as an estimator for the full sample predictor performance. The precise degree of this bias will depend on the ability of the learner to select a predictor that is arbitrarily close to the feature-label relationship present in $\mathbf{D}$. This can be precisely quantified using various results in statistical learning [109, 134], and this can even be used to construct confidence intervals for the full sample predictor performance based on the resubstitution estimate.

When the features of an item are represented by a sequence of d numerical values, the feature space $\mathbb{X}$ may be viewed as a subset of $\mathbb{R}^d$. The term d is called the **dimension** of the feature space. As discussed in chapter 2, the feature spaces in ANA often possess a very large number of dimensions relative to the number of items. This makes it easy for learners to select predictors with very high performance on the training set. In binary classification tasks, this means that there will always exist some selectable linear predictors that have perfect accuracy on the training set. As the consequence of this, the resubstitution based confidence intervals of statistical learning theory become too wide to be useful. Figure 3.2 illustrates this phenomenon in the case of linear, binary classification problems. Here, a linear predictor $u(\mathbf{D})$ is selected and then evaluated on a full sample of 400 items. The coloured lines illustrate the 95% confidence lower bound for the true predictor performance provided by the Vapnik-Chervonkis dimension [109] as a function of the observer resubstitution performance given by the x-axis and the black diagonal line. This bound is already very wide when $d = 1$, and it expands rapidly as d increases. By the time d reaches 17, a resubstitution accuracy of 100% only provides for a lower bound of roughly 10% to be expected on independent data. (The resubstitution bounds at the different dimensions may be compared to a similar bound based on a performance measurement produced on an independent test set; this is much narrower than all of them.) To estimate predictor (or learner) performance quantities in the high dimensional contexts typical of ANA, it is therefore necessary to use independent training and testing sets. That is, it is necessary to use cross validation.

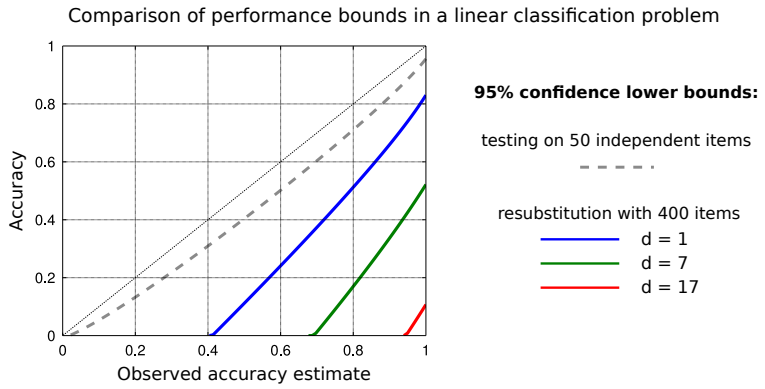

Figure 3.2: Comparison of one-sided confidence intervals for the predictor performance based on resubstitution or independent testing in a linear, binary classification task. The 95% Vapnik-Chervonenkis bounds implied at different feature space dimensions d are given for each possible observed resubstitution performance on a sample of 400 items. The bound implied by the same observed performance on an independent test set of 50 items is given for comparison. The distance of the lower bound from the diagonal provides a measure of the width of an interval.

3.3.2 Cross validation

Cross validation (CV) may be broadly defined as the use of separate training and testing sets to estimate performance [36]. This is achieved by producing one or more suitable train-test experiments $\mathbf{D} = \langle D_i \rangle_{i=1}^{l}$ by dividing it into disjoint parts. Different CV **strategies** do this in different ways. For most of the common CV strategies, each train-test experiment is created using two disjoint subsets of the integers $\{1,2,...,l\}$, denoted $I = \{\imath_i\}_{i=1}^{m}$ and $J = \{\imath'_i\}_{i=1}^{n}$, that have m and n elements respectively. By definition, $I \cap J = \varnothing$, and $m+n \leq l$.

Let $\imath_i$ denote the ith of the m indices contained in I. (The indices have some arbitrary order that allows them to be indexed themselves.) The training set $\mathbf{G} = \langle G_i \rangle_{1 \leq i \leq m}$ is defined

$$G_i := D_{\imath_i} \text{ for } 1 \leq m. \tag{3.9}$$

Similarly, where $\imath'_i$ is the ith element of the n elements of J, the testing set $\mathbf{H} = \langle H_i \rangle_{1 \leq i \leq n}$ is defined

$$H_i := D_{\imath'_i} \text{ for } 1 \leq n. \tag{3.10}$$

Most CV strategies place all items of $\mathbf{D}$ which are not used in $\mathbf{G}$ into $\mathbf{H}$. This means that J is defined by I as follows:

$$J = \{1,2,...,l\} \setminus I. \tag{3.11}$$

The result of the train-test experiment is the performance measurement $\gamma(u, \mathbf{G}, \mathbf{H})$ defined in equation (3.5).

CV strategies may comprise multiple train-test experiments. In this case, a CV strategy is

specified by a **block design** denoted $\mathbf{I} = \langle I_r \rangle_{r=1}^{R}$. This is a sequence of R index sequences, each specifying a training and testing set to be used in a **component train-test experiment**. The final performance measurement produced by a CV strategy is the mean performance observed over all component experiments. The validation function, Γ, takes a learner u, a block design $\mathbf{I}$, and a dataset $\mathbf{D}$ as its inputs, and returns the CV performance estimate. This may be defined as follows:

$$\Gamma(u, \mathbf{D}, \mathbf{I}) = \frac{1}{R} \sum_{r=1}^{R} \gamma(u, \mathbf{G}_r, \mathbf{H}_r), \text{ where} \tag{3.12}$$

$\mathbf{G}_r$ and $\mathbf{H}_r$ are the training and testing sets specified by I_r and $J_r = \{1, 2, ..., l\} \setminus I_r$ respectively. The difference in performance between two learners u and u' may be estimated by

$$\Gamma(u, \mathbf{D}, \mathbf{I}) - \Gamma(u', \mathbf{D}, \mathbf{I}). \tag{3.13}$$

The design $\mathbf{I}$ is itself a random variable. This means that, in addition to random variation associated with the generation of a dataset $\mathbf{D}$, a CV experiment has some internal random variation associated with the generation of the block design. Because no ordering of the items of $\mathbf{D}$ is used to assign them to different training and testing sets, *the distribution of* $\mathbf{I}$ *must be invariant to permutations of the indices* $\{1, 2, \cdots, l\}$ *for the items in* $\mathbf{D}$.

3.4 Common cross validation strategies

In this section, I shall outline the most common CV strategies.

3.4.1 Simple hold-out

The simplest CV strategy is the simple hold-out cross validation (SHOCV). A simple hold-out experiment entails a single train-test experiment parametrised by a randomly generated index subset I_1, typically selected from a uniform distribution of all $\binom{l}{m}$ m-size subsets of $\{1, 2, ..., l\}$.

3.4.2 Repeated hold-out

SHOCV uses only one of the $\binom{l}{m}$ equally train-test experiments of a given type. In repeated hold-out cross validation (RHOCV), E of these experiments are selected at random for their results to be combined. Thus, the I_r for $1 \leq e \leq E$ are all independently drawn from the set of $\binom{l}{m}$ m-size subsets of $\{1, 2, ..., l\}$. By combining the results of multiple train-test experiments, RHOCV is able to provide a lower variance estimator than SHOCV with the same choice of m.

3.4.3 Leave-p-out

As the E parameter of RHOCV approaches infinity, all possible train-test experiments appear an equal number of times. In this limit, the estimate of the CV experiment has no internal

randomness, and is invariant to permutations of items of **D**. In leave-p-out cross validation (LPOCV) this limit is reached more efficiently with a large design **I** whose elements are the $\binom{l}{m}$ distinct m-size subsets of $\{1,2,...,l\}$. When l is large and m far from 0 or l, $\binom{l}{m}$ is often too large for LPOCV to be computationally feasible.

3.4.4 K-fold

K-fold cross validation (KCV) uses K train-test experiments defined as follows. Let $\mathscr{A} = \{A_1, A_2, ..., A_K\}$ be a randomly generated partition of the indices into K disjoint subsets ($A_r \cap A_{r'} = \varnothing$ where $r \neq r'$). Where l is divisible by K, it is possible to enforce $|A_r| = l/K$ for all r. Otherwise, it may be necessary to select $\mathscr{A}$ such that

$$|A_r| = \begin{cases} \lfloor \frac{l}{K} \rfloor + 1 & \text{if } r \leq l \bmod K \\ \lfloor \frac{l}{K} \rfloor & \text{otherwise.} \end{cases} \tag{3.14}$$

The K index subsets I_r of KCV are specified

$$I_r = \{1,2,...,l\} \setminus A_r \,. \tag{3.15}$$

This means that

$$|I_r| = \begin{cases} l - \lfloor \frac{l}{K} \rfloor - 1 & \text{if } r \leq l \bmod K \\ l - \lfloor \frac{l}{K} \rfloor & \text{otherwise.} \end{cases} \tag{3.16}$$

By construction, KCV uses all the items of **D** for training and testing an equal number of times. The partition $\mathscr{A}$ may be generated using random permutations of the integers $\{1,2,...,l\}$.

3.4.5 Repeated K-fold

RHOCV can be viewed as combining E sequential SHOCV with different random instantiations of **I** to reduce the variance of the final performance estimate. In the same way, repeated K-fold cross validation (RKCV) combines E sequential KCV experiments. In RKCV, **I** comprises $R = EK$ index subsets. For e in $\{1,2,...,E\}$, index subsets $(e-1)K$ through eK are defined by a random partition $\mathscr{A}^{(e)} = \{A_1^{(e)}, A_2^{(e)}, ..., A_K^{(e)}\}$. Each $\mathscr{A}^{(e)}$ is independently and identically generated in the same way as $\mathscr{A}$ in KCV. That is, for $1 \leq r \leq EK$,

$$I_r = \{1,2,...,l\} \setminus A_{r \bmod K}^{(1+\lfloor r/K \rfloor)} \,. \tag{3.17}$$

3.4.6 Leave-one-out

Leave-one-out cross validation (LOOCV) uses l training sets of size m. These are specified by

$$I_r = \{1,2,...,l\} \setminus \{r\}\,, \text{ for } 1 \leq r \leq l\,. \tag{3.18}$$

LOOCV is a special case of KCV where $K = l$. It is also a special case of LPOCV (with $m = l - 1$). LOOCV may be attractive in situations where one wishes to use training sets that are as large as possible.

3.5 On the properties of cross validation strategies

This section describes how the expectation and the variance of a CV strategy are determined by the selection of a distribution for a block design $\mathbf{I}$.

3.5.1 Expectation

The expected value of the performance estimate produced by a CV experiment is the average of the expected values produced by all its component train-test experiments (corresponding to the $\mathbf{G}_r\ \mathbf{H}_r$ pairs).

In SHOCV, RHOCV, LPOCV, and LOOCV, the marginal distribution of I_r is the same for all r, and all train-test experiments are of the same type; they all entail training and testing sets with m and n items respectively. This is also true for KCV and RKCV when $l \bmod K = 0$. In all these cases,

$$\mathbb{E}\big[\Gamma(u,\mathbf{I},\mathbf{D})\big] = \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\big[\gamma(u,\mathbf{G}_r,\mathbf{H}_r)\big] \tag{3.19}$$

$$= \mathbb{E}\big[\gamma(u,\mathbf{G}_r,\mathbf{H}_r)\big]\,, \tag{3.20}$$

for any $1 \leq r \leq R$, because $\mathbb{E}[\gamma(u,\mathbf{G}_r,\mathbf{H}_r)]$ is the same for all r. That is, the expected value of the final CV performance estimate is the same as the value expected in any one of its component experiments: the learner performance associated with a training set of m items. CV strategies in which the I_r have the same marginal distribution may called **commensurate**.

In KCV and RKCV more generally, $(l \bmod K)$ out of K train-test experiments have a training set of size $m_1 = l - \lfloor l/K \rfloor - 1$, while the remainder have a training set of size $m_2 = l - \lfloor l/K \rfloor$. This means that the expected value of the experiment may be intermediate between the learner performances associated with the training set sizes m_1 and m_2. In practice, the difference between m_1 and m_2 may be negligible, and KCV or RKCV may be taken as an approximately unbiased estimator of the learner performance associated with either.

3.5.1.1 Training set bias

The bias of a CV performance estimator is defined by the expectation of the estimator and the quantity it is being used to estimate.

The full sample predictor performance. The full sample predictor performance varies randomly with the dataset $\mathbf{D}$. Its expectation is the learner performance associated with a training set of size l. Because the expectation of $\Gamma(u,\mathbf{D},\mathbf{I})$ is the learner performance associated with the lower training set size $m < l$, it is likely to be lower. This means that, when taken as an estimator of the full sample predictor performance, a CV performance estimate is pessimistically biased. The degree of its bias will be dependent on the difference between m and l, with larger m producing smaller levels of bias. It is for this reason that LOOCV (where $m = l-1$) is often spoken of as 'having low bias'.

A learner performance. As described in section 3.2.2, learner performances are a function of training set size. Where a CV experiment with training set sizes of m is used to estimate a learner performance associated with a size of m', it will provide an unbiased estimator when $m = m'$. If $m \neq m'$, then the CV estimator will have some bias determined by the difference between performance of the learner at the two training set sizes. In general, higher training set sizes lead to higher learner performances, so the bias will be negative when $m < m'$.

3.5.2 Variance

For a given u, the outcome $\Gamma(u,\mathbf{D},\mathbf{I})$ of a CV experiment is determined solely by the value of the random variables $\mathbf{D}$ and $\mathbf{I}$. As pointed out by Fuchs et al. [135], by using the law of total variance, one may decompose its variance as follows:

$$\mathbb{V}\mathrm{ar}\left[\Gamma(u,\mathbf{D},\mathbf{I})\right] = \mathbb{E}_{\mathbf{D}}\left[\mathbb{V}\mathrm{ar}_{\mathbf{I}}\left[\Gamma(u,\mathbf{D},\mathbf{I})|\mathbf{D}\right]\right] + \mathbb{V}\mathrm{ar}_{\mathbf{D}}\left[\mathbb{E}_{\mathbf{I}}\left[\Gamma(u,\mathbf{D},\mathbf{I})|\mathbf{D}\right]\right]. \tag{3.21}$$

The term $\mathbb{E}_{\mathbf{I}}\left[\Gamma(u,\mathbf{D},\mathbf{I})|\mathbf{D}\right]$ appearing in the second right summand of equation (3.21) is the average taken over all possible values of $\mathbf{I}$. For all the strategies of constant training set size discussed in section 3.4, each of the I_r has a uniform marginal distribution over the set of all m-size subsets of $\{1,2,...,l\}$. This means that, when the expectation is taken over $\mathbf{I}$, the result is the average performance obtained in all $m-n$ size train-test splits of $\mathbf{D}$. This is simply the result of LPOCV on $\mathbf{D}$. Accordingly, $\mathbb{V}\mathrm{ar}_{\mathbf{D}}\left[\mathbb{E}_{\mathbf{I}}[\Gamma(u,\mathbf{D},\mathbf{I})|\mathbf{D}]\right]$ is the variance of LPOCV. This is the *irreducible*, or 'external', component of the variance of a CV experiment. It is the same for all strategies where the I_r have a given uniform marginal distribution, regardless of their number and how they are generated (e.g., KCV, RKCV SHOCV and LPOCV using m out of l items for training).

The first right summand of equation (3.21), that is to say $\mathbb{E}_{\mathbf{D}}\big[\mathbb{V}\text{ar}_{\mathbf{I}}[\Gamma(u,\mathbf{D},\mathbf{I})|\mathbf{D}]\big]$, is the variance conditional on $\mathbf{D}$. It is this term that may differ between commensurate strategies with the same expectation. Because the results of sequential CV experiments using E different random instantiations of $\mathbf{I}$ are independent conditional on $\mathbf{D}$, this part of the variance is proportional to E^{-1} in RKCV and RHOCV (where E denotes the number of experiment repetitions). This is the *reducible*, or 'internal', component of the variance of the CV experiment.

Where two CV strategies have the same expectation and use the same number of train-test experiments but one of them has a lower variance, the lower variance strategy may be called more **efficient**. This is because it will provide a more precise estimator of a learner performance while using the same amount of computational resources. Where two commensurate strategies have different variances, the more efficient one has a lower reducible component.

3.6 Item subpopulations and stratification

In many contexts, a population of items may be divided into two or more disjoint subpopulations. These subpopulations may be defined by different label values, different feature values, or different covariates associated with the items. In supervised learning, **stratification** is the practice of producing testing and/or training sets in a way that ensures they have a fixed fraction of items from each subpopulation. It is seen very often in classification contexts where the categorical labels provide a natural and important division of the items. In ANA, one might use either diagnosis (label), hippocampal volume (covariate/feature) or age (covariate/feature) for stratification.

One normally regards the l items of the dataset $\mathbf{D}$ as being i.i.d. from a single population. In stratification, one divides them into S groups corresponding to some S subpopulations. The full number of items l may be written as the sum $\sum_{s=1}^{S} l_s$, where l_s represents the number of items from subpopulation s. While all items are independent, only those within a single subpopulation are identically distributed. In a non-stratified context, one produces a train-test split with m items in the training set using a random index set I_r that is uniformly distributed over all m-size subsets of the indices $\{1,2,...,l\}$. In a stratified context, the training set must contain a fixed $m_s \leq l_s$ items from subpopulation s. These are selected with a random m_s-size index subset of the l_s indices corresponding to items from subpopulation s. After subsets are taken from each of the subpopulations, the final training index set I_r is the union of those from each subpopulation. In this context, the distribution of I_r is required only to be invariance to permutation of the indices $\{1,2,\cdots,l\}$ which preserve the labels of the items in $\mathbf{D}$. This produces training and testing sets whose items are independent, but can only be taken as i.i.d. within a

single subpopulation.

The subpopulation training set contributions m_s may be chosen to ensure that each training set is representative of the full sample. This should mean that m_s/l_s should be roughly the same for all s. In KCV and RKCV, one may divide the items of each subset with a partition in the same way as one did the full set of items in the non-stratified cases. One then combines the resulting subpopulations into a new partition whose ith index subset is the union of the ith index subsets from all the individual subpopulation partitions.

3.6.1 Role of subpopulation composition in defining performance quantities

In a non-stratified context, the performance of a fixed predictor is the expectation of the result produced in a testing experiment in which an arbitrary number of items are used in a testing set; because all items are drawn from the same population, the expected performance on any one of them is the same. In a stratified context, a predictor will have different **subpopulation performances** on the different subpopulations. The performance of the predictor on subpopulation s is the performance expected in a testing experiment in which all items in the testing set come from that subpopulation. In binary classification tasks where one class is associated with health and the other with disease, the population performance associated with the health class is called **specificity**, while the performance associated with the disease class is called **sensitivity**. In a general testing experiment for a fixed predictor in which the testing set comprises items from multiple subpopulations, the performance expected will be an average of the performances associated with each subpopulation weighted according to their relative contributions to the testing set.

In the same way as a predictor, a learner also has a performance associated with each subpopulation. This is the expected result of a train-test experiment in which the testing set has items from that subpopulation alone. In a non-stratified context, the performance of a learner is defined only by the total number of items in a training set. In a stratified context, the precise number of items from each subpopulation will determine the distribution of predictors produced. This will determine the subpopulation performances which in turn determine the full population performance. Thus, the expected performance of a learner in a train-test experiment depends both on the composition of the training set and that of the testing set. In a stratified context, commensurate CV strategies will have training sets with a constant size and subpopulation composition.

3.7 Dependencies between component cross validation results

This section provides a brief outline of the problem of dependency in CV results. This topic is dealt with in more detail in chapter 8, but it is introduced here as it is important for understanding some of the validation choices made by ANA researchers.

A general CV experiment may be contrasted with a simple testing experiment with a *fixed predictor t* and a testing set **H** of i.i.d. items. In the simple testing experiment, the performance results $Q_i = \phi\left(Y_i, t(X_i)\right)$ where $H_i = (X_i, Y_i)$ are i.i.d. random variables. This allows standard statistical procedures to be employed to produce confidence intervals or tests for their shared expectation, the predictor performance μ_t.

In a train-test experiment with a pre-specified learner u and a random training set **G**, the predictor $T = u(\mathbf{G})$ is now a random variable. The performance results $Q_i - \phi\left(Y_i, T(X_i)\right)$ are only independent *conditional upon* T. The expectation of the Q_i conditional on T is the random variable M_T, the predictor performance of T. The marginal expectation of M_T, and thus of Q_i, is the learner performance μ_u. When T happens to be a better than average predictor, M_T is higher, and all the Q_i are likely to be higher. This shared dependence on T/M_T means that the Q_i are marginally dependent with some positive correlation.

In order to produce confidence intervals or tests for the learner performance μ_u in a train-test experiment, one may use a **fixed predictor model** in which the Q_i are treated as if arising from a fixed predictor t such that $\mu_t = \mu_u$. This assumes the Q_i are independent and neglects the dependency between them due to the variability of M_T. As a consequence, the resulting intervals and tests for the learner performance are not strictly valid, and will have worse behaviours than expected. This is known as the **problem of dependency**.

In a more general CV strategy where multiple train-test experiments are performed using different partitions of a single dataset **D**, the results of the component train-test experiments are dependent through their shared use of the data for training and testing. *The greatest degree of correlation will be seen in strategies where items are repeatedly reused for testing, such as RHOCV and RKCV*. While KCV does not reuse items for testing, there is still some correlation between the results of its component train-test experiments (even where $K = 2$), so they cannot be treated as independent [38,40,135]. As in the single train-test experiment case of the previous paragraph, the lack of independence between component results can undermine the validity of conventional statistical procedures.

3.8 Cross validation in AD ANA

This section reviews the role of CV in AD ANA, the types of strategies used, and the size and composition of the training sets.

3.8.1 What is CV being used to estimate?

The vast majority of AD ANA studies advance new learner specifications, or a new step for constructing learner specifications, on the basis of the superior learner performance those specifications can offer over alternatives in a particular task [7, 29]. While all ANA studies use CV to estimate performance, it is not often explicitly stated which performance quantity is intended as the figure of merit. It is possible that this may differ between studies. When only a fixed predictor is constructed and evaluated, it may feasibly be the performance of the specific predictor constructed. More generally, however, multiple train-test experiments are used to evaluate performance, and changes in performance are explained as being due to changes to a learner specification, without reference to the particular training sample used. For this reason, for the remainder of this thesis, *I shall take it that the following are key goals of AD ANA research: 1) the identification of learners with superior performance and 2) the estimation of those learner's performances*. This should allow the field to identify methods with the highest expected performance in some imagined future application with an as yet unseen training set of items similar to those seen in research. It should also allow one to judge the level of benefit expected with the introduction of those methods, something that is crucial when deciding whether or not research methods should actually be applied.

The future application of AD ANA methods could be pre-selection or response tracking for a drug trial, or in a clinic supporting routine medical decisions. The precise details of the future training set are unknown, so there is no precise training set size (or subpopulation composition) to determine the learner performance being estimated. Rather, there is some anticipated range of sizes. It is assumed that learner performance ranking will be stable over this range, so one can reliably rate one learner as better than another. Ideal cross validation strategies will estimate learner performances associated with training sets in this range with high precision and low bias.

3.8.2 Training set sizes and subpopulation compositions

The sample sizes used in AD ANA cover a range of less than 50 to over 1500, though the large majority of studies have a sample of less than 300 subjects [24, 25, 29]. Accordingly, there is a very large range of training set sizes considered. (One should note that not all items in the total sample whose size is reported in a study are reported in all its CV experiments,

so the typical training set may be appreciably smaller than the typical sample.) Even when researchers use the same dataset (e.g., ADNI), different data requirements mean that different studies will have different numbers of subjects. In particular, studies using sMRI will tend to have the largest samples, while those rarer modalities or multi-modal imaging will tend to have the smallest [29]. As well as changes in the size of the sample, in classification studies, there are changes in its class composition. It is very rare to find a pair of studies using precisely the same training set size and composition (see [7, 24, 29]).

3.8.2.1 Issue of disparity

As others have pointed out [29, 136], the variability of training set sizes, training set subgroup compositions and dataset origins seen in ANA research makes it difficult to compare learners across multiple studies. Though methods are often compared on the basis of performance estimates derived from different studies, this comparison is not always meaningful; while learner A may appear to have higher performance than learner B, if learner A was evaluated on a larger or more balanced training set, then it might have been the case that learner B would have outperformed it under identical conditions. This problem can only be overcome by comparing learners side-by-side in the same experiment. This does not occur often, as the vast majority of studies are interested in demonstrating the superiority of their authors' novel method over some particular alternative [7, 29] rather than comparing options from the expansive set of possible methods.

The lack of controlled comparison of learner performances is a great impairment to what should be a key goal of ANA research: the identification of the best learner for the task. Even leaving aside the issue of selection bias, this provides a strong motivation for comparative studies in which a variety of methods are compared under controlled conditions [22, 88]. A notable effort to improve standardisation came from the authors of [136], who published lists of subject ID's to allow for standardised SHOCV and KCV experiments.

The problem of disparity becomes even greater if one wishes to identify which *parts* of learner specifications lead to higher performance, as any two ANA studies are unlikely to implement even one component of their learners in precisely the same way.

3.8.3 Cross validation strategies

The majority of AD ANA studies use a single repetition of KCV to estimate performance [49, 50, 105, 137], with the common choice of K being 10 [24]. LOOCV is not uncommon, particularly in small samples [96, 138–140]. RKCV is often used to improve precision [46, 47, 141]. In other cases, SHOCV is used [28, 88, 96]. This may be for the sake of

simplicity, or to avoid the issues of dependency introduced in section 3.7. SHOCV also offers a very natural way to implement prediction **challenges**, where contestants must submit predictions for a testing set for which no labels are provided [25]. These are important as they avoid bias due to any deliberate or unwitting optimisation on the testing set by competing researchers (an issue that is discussed in more detail in chapter 4). There are a few examples of RHOCV; where it occurs, it appears to have been motivated by the desire to produce new statistical procedures for the treatment of results [45, 102]. With the majority of AD ANA studies involving classification, class stratification is very common with every strategy but LOOCV.

Summary

In this chapter, I have introduced the key concepts of supervised learning. These include the key performance quantities that one might wish to estimate, and the experiments that one might use to measure them. I have described CV experiments for performance estimation, and explained why they are necessary. I have introduced two important issues related to CV that will be revisited later in this thesis: the problem of dependency between component CV results and the problem of over-fitting and selection bias. I have reviewed the role of CV performance estimation in AD ANA and described its practice. The rest of this thesis will be dedicated to searching for improvements to that practice.

Part II

Bias

Chapter 4

Bias in published performance results

This chapter will review potential sources of bias in the published performance results of the AD ANA literature. Its purpose is to provide the necessary background information for the empirical study of chapter 5. This chapter has a particular focus on bias due to the selective reporting of results, as that is the source considered in the empirical study. However, I shall also provide a brief description of all sources of bias to allow them to be clearly differentiated.

4.1 Significance of bias in performance measurement

Before any ANA methods can be applied for clinical trial enrichment or decision making, it must be clear that they offer better outcomes than the existing alternatives. If ANA methods are brought into the clinic on the basis of optimistically biased performance estimates (e.g., overestimate diagnostic accuracy), their introduction may actually lead to clinical decisions that are *worse* than was previously the case. Methods introduced on the basis of overestimated performance may actually hinder the discovery of new therapies in clinical trials, and they can cause harm to patients though a poorer selection of treatment. The reduction of bias is therefore an important concern.

4.2 Sources of bias unrelated to selection

This section describes sources of bias unrelated to selection to allow them to be distinguished.

4.2.1 Bias due to population shift

Population shift or concept drift occurs when a population of items to which a predictor or learner is applied changes. As one would expect, a change in population leads to a change in learner or predictor performance. This is a problem most often considered in online learning problems, where training items are made sequentially over time and a predictor must be regularly updated. In the context of AD ANA, clinical practice must contend with scanner types and patient populations that may be unlike those in research settings. When research methods

are applied to these new settings, performance results may fail to generalise [42]. The bias of population shift is associated with learners being unrepresentatively well tailored to a particular *population*.

Various works in the literature have considered the problem of population shift [22, 43, 142]. In particular, Abdulkadir et al. have pointed out how including multiple scanner types in a training sample can improve generalisation without necessarily harming the performance on any one type [43]. Similarly, one solution to the problem of population shift may simply be to use a diverse dataset (in terms of subject population, scanner type, etc...) for predictor training and evaluation.

4.2.2 Bias due to differences in training set size

As described in section 3, the learner performance is a function of the training set size and composition. Different training set sizes and compositions provide different levels of information about the full distribution of the items. If a study considers itself to be evaluating learners to be trained on a random training set of a larger size than that used in CV, then CV will be a biased estimator of the relevant learner performance. Similarly, if a study imagines that the learners it evaluates will be trained for general use on the full available sample used, then the CV performance estimates are also likely to be downwardly biased (as larger training sets tend to produce better predictors). The bias resulting from differences between the evaluated and the imagined training set sizes may be called **training set bias**.

Training set bias is unavoidable, but it may not always be important. There are three reasons for this. The first of these is that diminishing improvements in learner performance with increasing sample size may mean that, once a 'reasonable' training set size is achieved, this bias should be relatively small. The diminishing improvement in AD classification tasks are apparent in the performance-sample sizes response curves in the feature selection study of Chu et al. [105], as well as in my own variance study in 2014 [35]. Figure 4.1 presents a result from the latter, describing the relationship between expected classification accuracy and sample sizes in a balanced sample.

The second reason that training set bias may be unimportant is the stability of learner *rankings* with training set size. While learner performances may increase with training set size m, they all increase together. This means that differences between learner performance change more slowly with m than do the performances themselves. This phenomenon means that training set bias is often less important than anticipated in learner selection, where variance plays an important role [143]. In AD ANA, differences in performance between learners are often explained in terms of general principals that do not imply an important role for sample

size. (For instance, a preferred learner may be said to exploit an additional source of information or use a particular algorithm that is more suitable for the context in question.) This makes the stability of rankings over m seem particularly plausible.

The third and final reason that training set bias may be less of a concern than the other types is that will typically be *pessimistic* rather than *optimistic*, as future applications of research methods may be expected to have access to larger datasets than those used in initial studies. No harm is likely to result if an applied research method has a higher performance than anticipated.

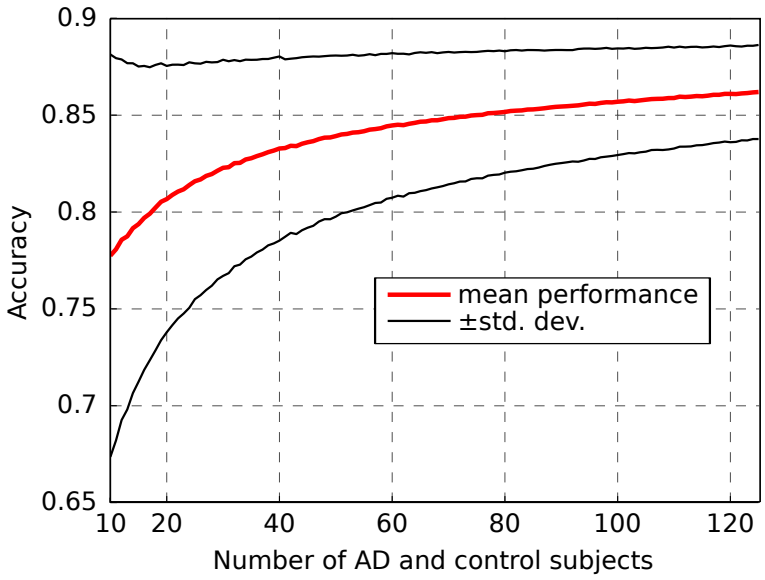

Figure 4.1: The average accuracy observed in the discrimination of AD subjects and healthy controls from the ADNI study when different sample sizes are used in KCV with $K = 10$. These results were measured using experiments on many class-balanced subsets of larger sample. The number of subjects from each class in the training set is approximately 0.9 times the number of those subjects in the subset as marked on the X-axis. The diminishing improvement associated with increased training set size can be clearly seen.

The simplest defence against training set bias is to use as large a sample as possible; at larger samples, the diminishing returns phenomenon means that fractional changes in training set size will have less of an effect of learner performance. Additionally, one may use one of several stratification-like strategies discussed in 6.3 to ensure that the training sets produced in a CV strategy are more representative than they would be if produced using purely random partitions.

4.3 Selection bias

On the main goals of AD ANA research is the development of learners with superior performance. The majority of studies are based around the development and evaluation of new

learner specifications that are justified primarily on the basis of superior performance. The improvement in performance may be measured relative to some baseline method, or may be claimed over *all* alternative specifications [29]. Already, a vast number of learner specification have been considered and evaluated, particularly for classification. In the supplementary document [7], I have compiled a list of 470 unique publications detailing new learners for AD classification tasks. This list is non-exhaustive, but its length already exceeds the size of a typical sample (less than 300 [29]).

An inevitable consequence of the desire for higher performance is that performance results will be reported selectively. Studies on a new learner that yields unimpressive results are unlikely to be accepted for publication. Groups are unlikely to even submit for publication any new methods that appear unfavourable, and will instead focus their attention on refining methods that do well in preliminary investigations. Because the performance estimates that are reported have been **selected** on the basis of their relatively high values, they will become biased estimators of their own expectations. This type of bias, which I call **selection bias**, is simply the manifestation of publication bias in performance estimation. Selection bias may also be considered a form of regression to the mean [144], defined as the phenomenon where the extreme measurements of a set of variables are, on average, found to return towards normal levels on repeated measurement. Here, the random measurements are performance estimates for some set of learners, and the highest of these move down towards the mean when replicated in an independent experiment.

While selection bias could be introduced by only reporting performance estimates that reach some fixed required threshold, a better model for AD classification would be the selective reporting of performance estimates conditional on their *relative* values. This is because selection will often occur after the comparison of multiple candidate learners on identical or overlapping datasets. The practice of using CV on the full sample to identify the learner parameters (e.g., the C parameter of the SVM algorithm) that lead to the highest estimated performance and then reporting the highest performance estimates, also called 'double-dipping', has long been recognised as unacceptable practice in neuroimaging applications [23, 145].

Though straightforward parameter selection is now widely regarded as bad practice, it is less well known that the entire specification of a learner may be regarded as a parameter. The precise specification of the image processing, feature selection and learning algorithm used in a learner may all be varied to affect performance. When one compares multiple learners in an experiment and then reports only the performance of the best, this is precisely analogous to naked parameter 'double-dipping' and is liable to cause selection bias. This has been previously

noted in an imaging context by Rao et al., who observed that their best performing learners suffered a significant drop in performance when applied to independent data [146].

At the heart of the selection bias issue is a very simple principle: when a number of quantities are estimated and ranked in the presence of measurement noise, the estimates associated with different rank positions can become biased. There is a sort of uncertainty principle at play: one cannot simultaneously identify which quantity is the highest, and estimate what that quantity is without bias. In ANA research, one cannot use the same dataset to identify the best performing of two or more learners and at the same time provide an unbiased estimate for the performance of the one identified as the best.

Selection bias may be contrasted with bias due to population shift. While bias due to population shift is due to a learner being unrepresentatively well tailored to a particular *population*, selection bias is due to a learner being unrepresentatively well tailored to a particular *sample or CV experiment*.

4.3.1 How selection occurs in ANA

The process of selection is most obvious in a journal that rejects publications detailing new methods without promising results and in papers that evaluate multiple methods and then report only the estimates associated with those that are apparently the best. Selection bias may also occur in another, less transparent way.

While published works proposing new ANA methods often speak as if each new learner was fully specified in complete isolation from the data, this is often likely to be a fiction; in practice, researchers draw on previous experiences with the few datasets available for the learning task under study, and abandon projects that yield unfavourable results. This means that, even before any selective publication, it is likely that methods submitted to journals represent the best of multiple attempts. That is to say, even before any screening by reviewers, the associated results may already be the result of a selection process that can introduce bias. This situation is analogous to that seen in clinical research, where the most common reason that non-significant findings go unpublished is that they are never submitted for publication in the first instance [147].

4.4 A simple model for selection bias

A series of quantities $\langle \mu_i \rangle_{1 \leq i \leq n} \in \mathbb{R}^n$, have fixed but unknown values that are estimated by a corresponding sequence of unbiased estimators $\langle X_i \rangle_{1 \leq i \leq n} \in \mathbb{R}^n$ associated with some measurement

experiment. The estimators are unbiased, so for all i,

$$\mathbb{E}[X_i] = \mu_i .$$

The rank of the ith measure in the full series of n is given

$$R_i = \sum_{j=1}^{n} \mathbf{1}_{X_i \leq X_j} . \tag{4.1}$$

The expectation of a measure conditional on it attaining the lowest rank may be written

$$\mathbb{E}[X_i | R_i = 1] > \mu_i .$$

Where $X_{(k)}$ is the measurement taking the kth rank in a given observed sequence, the expected value associated with the kth rank is given

$$\mathbb{E}[X_{(k)}] = \sum_{i=1}^{n} P(R_i = k) \cdot \mathbb{E}[X_i | R_i = k] .$$

Where $\mu_{(k)}$ is the true quantity corresponding to $X_{(k)}$ (i.e. μ_j such that $R_j = k$), the selection bias associated with the kth position may be defined as

$$\mathbb{E}[X_{(k)} - \mu_{(k)}] = \sum_{i=1}^{n} P(R_i = k) \Big(\mathbb{E}[X_i | R_i = k] - \mu_i \Big) .$$

This is simply the expected difference between the measurement used to select a quantity based on its rank and another measurement of the same quantity in an independent experiment. Because performance estimation is unbiased, this is the expected random effect associated with the kth ranking measurement.

Where measurements are identified with performance, high values are considered desirable, and the selection bias associated with the highest ranking measurement is the average disappointment where one expects the highest performance in an initial set of measurement to be replicated in a second. The quantities μ_i could be the true performances of learners, making the quantities X_i measurements of these produced with CV. Alternatively, the quantities μ_i could be the 'true expected performances' of professional baseball players, making the measurements X_i the observed performances of those players in a particular season. It has been observed that baseball players who do exceptionally well in a particular season typically do not do so well in the next [144]. In this context, the selection bias associated with the highest rank is the expected drop in observed performance of the highest scorer between the first and the subsequent

seasons.

An illustration of the phenomenon of selection bias, along with an explanation in the context of learner selection, is presented in figure 4.2.

The degree of selection bias is dependent on the degree of the 'noise' and its contribution to the rankings of the measurements. When the differences between the true quantities are sufficiently great relative to the measurement noise, then the ranking of the measurements is essentially fixed. This means that the kth ranking measurement is invariably associated with the kth ranking quantity, μ_k. As demonstrated below, this makes selection bias equal to zero by definition;

$$\mathbb{E}[X_{(k)} - \mu_{(k)}] = \mathbb{E}[X_k - \mu_k] = 0\,.$$

4.4.1 Relationship with determining factors in a simple model

Using the simple model of equally spaced quantities between -1 and 1 measured with a fixed amount of Gaussian noise, I have performed a simple computational experiment to illustrate the importance of the noise level and number of quantities in determining bias. For noise levels in the set $\{0,1,2,3,4\}$ and quantity numbers in the set $\{1,4,16,64,256\}$, I used 10000 simulated measurement experiments to calculate the bias associated with the highest measurement. These are presented in figure 4.3. It can be seen that bias is roughly proportional to noise, and that it increases 'sub-logarithmically' with the number of quantities.

4.5 The dangers of shared data

One straightforward way to guard against selection bias is through independent repetition of measurement experiments. Biased results and the resulting spurious conclusions can be identified by their inconsistency with independent measurements. Unfortunately, this is not often possible in AD ANA research. The field relies on collections of imaging and clinical data that are large and expensive. For individual studies to reproduce these solely for their own use would be both impractical and immensely wasteful. While large data sharing projects such as Alzheimer's disease neuroimaging initiative (ADNI) have meant that a much greater number of researchers have access to the sample sizes required to robustly measure learner performances, it also means that some of the random effects (or "noise") in performance measurement are shared between studies. Naturally, the greater the overlap between two studies' samples the more correlated their performance estimates will be. While published studies sharing data are not truly independent, they may have the appearance of being so. Without careful interpretation, the results of one may be taken as confirmation of the other, giving spurious effects the illusion of repeatability.

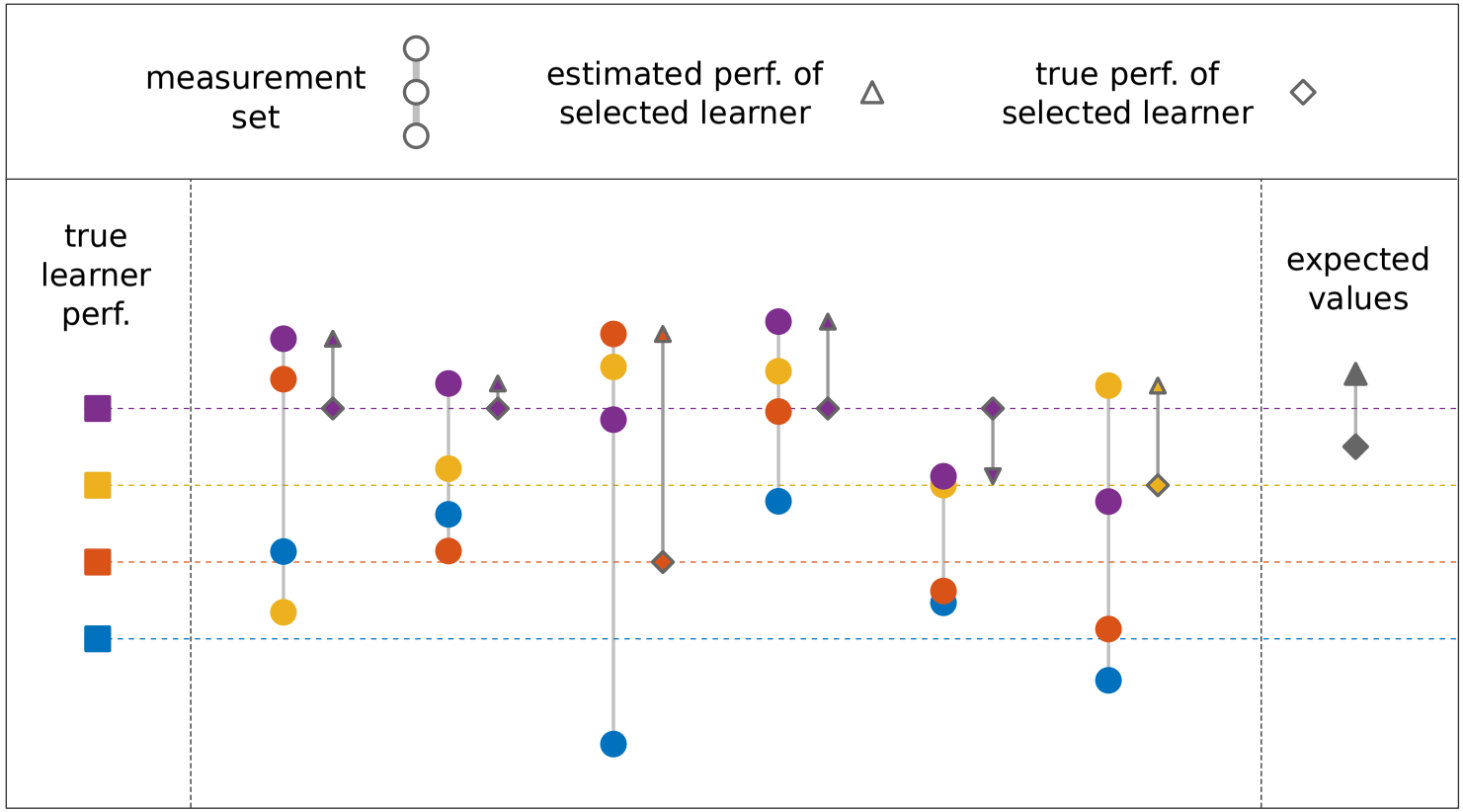

Figure 4.2: An illustration of selection bias. The fours squares on the left of the diagram represent the true performance of four learners. The six vertical bars in the centre represent independent sets of unbiased performance measurements produced using cross validation. Each circle represents estimated performance of the learner of the same colour. The triangles beside the bars denote the highest observed performance measurement corresponding to the learner identified as the best. The diamonds represent the true performance of the learner selected as the best. The displacement between the diamond and the triangle represents the random effect associated the highest performance measurement. On the right are the expected values of the estimated and true performances of the learner selected as the best. The difference between these is the expected bias of the highest measurement in a given experiment

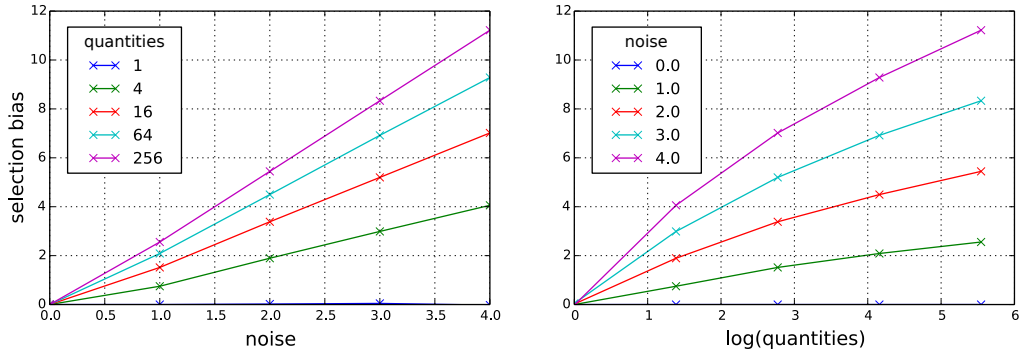

Figure 4.3: Effect of quantity number and noise on selection bias associated with the maximum measure $X_{(1)}$.

At the current time, the majority of AD ANA research involves the evaluation of numerous AD classification pipelines on the shared dataset provided by ADNI in the search for the best [24,29]. As the research community searches through the space of possible learners and selects the best of them for publication based on CV in the shared dataset, it imitates the behaviour leading to bias in the work of a single group. This leads to what is effectively a distributed selection process that, like any other selection process, will be liable to selection bias. Due to the great research interest created by AD's societal importance, many learners have already been validated and considered (including the 470 described in the supplementary material [7]). Though a wide variety of methodological options (pre-processing, feature selection, classification algorithm) provides more scope for real improvement, it also creates more opportunity for selection bias.

The problem of 'collaborative over-fitting' (where over-fitting is used the colloquial sense to denote selection bias) has been observed in popular machine learning benchmark datasets such as MNIST [148, section 3]. It is also evident in various other machine learning competitions. For instance, several contestants at Kaggle[1] have been able to predict which passengers would survive the Titanic disaster of 1912 with 100% accuracy based on ticket and demographic information alone. Clearly, no perfectly successful prediction rule should exist for such an outcome.

4.6 Selection bias and over-fitting in model selection

The term over-fitting is often used colloquially to describe selection bias in machine learning contexts [148], but over-fitting and selection bias are actually distinct concepts. This section will provide a precise definition of true over-fitting and selection bias as they appear in two analogous model selection tasks:

[1] https://www.kaggle.com/c/titanic/leaderboard

- the selection of a predictor by a learner to maximize predictive performance as estimated using resubstitution on a training set or
- the selection of a learner to maximise performance as estimated using CV in an available sample.

The first task is the most widely known, but the two are very similar. Both tasks involve some noisy **in-sample** performance estimate derived from a limited sample being used to rank a series of potential **models** (predictors/learners). In both tasks, the true goal is the selection of the models with the highest true **out-of-sample** performance, defined as that expected on independent data. An important consideration in both tasks is the **model complexity**, defined as the size or diversity of the set of models from which a high performing candidate is to be selected. The key correspondences between abstract and concrete terms in both tasks are illustrated in table 4.1.

	predictor selection	learner selection
in-sample performance estimate	resubstitution	CV
out-of sample performance	full dataset predictor	learner
model complexity	diversity of predictors	diversity of learners

Table 4.1: Meaning of model selection terms in predictor and learner selection

In abstract terms, true over-fitting occurs when increasing model complexity causes the out-of-sample performance of the selected model to decrease. In this situation, additional effort spent on model optimisation may actually be counterproductive. Selection bias occurs when, due its use in the selection of a model, the in-sample performance estimate is an optimistically biased estimator of the true out-of sample performance. *While selection bias will occur in almost all model selection contexts, true over-fitting only occurs when model complexity is excessive.* Over-fitting cannot occur in the absence of selection bias, but selection bias can occur in the absence of over-fitting.

In order to make the key terms a little more clear, I shall now provide a more concrete description of over-fitting and selection bias in both model fitting tasks.

4.6.1 Predictor selection

Recall that $g(t,\mathbf{H})$ represents the performance evaluation of a predictor t on a testing set $\mathbf{H}$. In predictor selection by a learner u on a dataset $\mathbf{D}$, the resubstitution performance $g(u(\mathbf{D}),D)$ is the in-sample performance estimate. The out-of-sample performance is the full sample predictor performance $\mathbb{E}_{\mathbf{H}'}[g(u(\mathbf{D}),\mathbf{H}')]$, where $\mathbf{H}'$ is some independent set of i.i.d. items from the same population as those in $\mathbf{D}$. As discussed in section 3.3.1, predictors selected on a train-

ing set acquire a special relationship with its items, making them no longer representative for the purpose of performance estimation. This is the manifestation of selection bias in predictor selection.

In predictor selection, model complexity is the 'flexibility' or diversity of a predictor set from which a learner must make a selection so as to maximise the resubstitution performance. (Depending on the prediction task, this may be quantified through the Rademacher complexity [134] or Vapnik-Chervonenkis dimension [109]). The greater this is, the greater the potential for over-fitting and selection bias.

A good illustration of the phenomena of selection bias and over-fitting is provided by polynomial regression. This can be regarded as a prediction task in which the items are the feature-label pairs (X,Y) where $X,Y \in \mathbb{R}$. In this context, predictors correspond to curves assigning each $x \in \mathbb{R}$ to a unique y. Let u_w denote the learner which selects a w-degree polynomial predictor on $\mathbf{D}$ so as to maximise the mean negative square error (i.e., to minimise the mean squared error). Higher values of w correspond to higher flexibility.

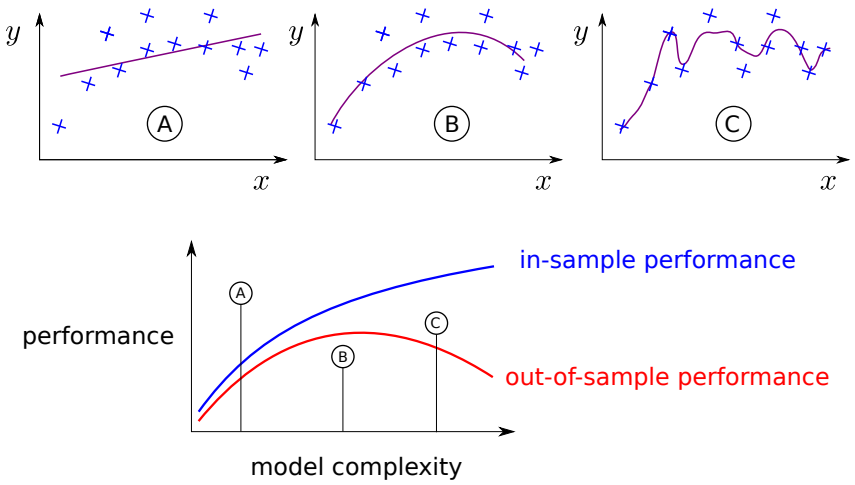

Figure 4.4: Fitting and over-fitting in a simple regression problem. Above, polynomial fits using three ascending levels of complexity: A, B and C. Below, illustration of the relationship between model (predictor) complexity and performance as measured on the training set or an independent sample. As model complexity increases, the error on the training set decreases. So does the error on independent testing data, but only up to a point.

The effect of w is illustrated in figure 4.4. When w is very low (A), the learner has little flexibility. The predictor will have a low resubstitution performance and a low true predictor performance, but selection bias will be low. As w increase to an optimum value (at B), both the resubstitution and true predictor performance increase as the predictor improves, though they begin to diverge as selection bias increases. As w moves beyond this optimum point, $u_w(\mathbf{D})$

begins to reflect more of the 'noise' variability in **D**. Though the resubstitution performance will only continue to increase, the true predictor performance will be reduced.

4.6.2 Learner selection

Consider a set of learner candidates $\mathbb{U}' = \{u_1, u_2, ..., u_k\}$. The set could contain learners with entirely different specifications (with feature extraction, learning algorithm, etc...), or could simply contain the learners corresponding to the different parameters settings of a singular learning algorithm (e.g., different values of K for the K-nearest neighbours algorithm). To select the learner with the highest performance, one may evaluate all possible selections using CV with block design **I** on a dataset **D**. After this, one then selects the learner with the highest performance estimate. This produces a selection u where

$$u = \arg\max_{u' \in \mathbb{U}'} \Gamma(u', \mathbf{D}, \mathbf{I}), \tag{4.2}$$

where Γ represents the validation function that takes a learner u, a block design **I**, and a dataset **D** as its inputs, and returns the CV performance estimate. In this case, because **D** has been used to select u, the performance estimate $\Gamma(u, \mathbf{D}, \mathbf{I})$ is now subject to selection bias, making it an overly optimistic estimator of the learner performance $\mathbb{E}_{\mathbf{G},\mathbf{H}}\big[\gamma(u, \mathbf{G}, \mathbf{H})\big]$, where **G** and **H** denote training and testing sets of independently generate items. In other words, u is peculiarly well suited to predicting the labels of **D** in a CV experiment, because it has been selected for that property.

Though it is not commonly discussed, true over-fitting in the context of learner selection by CV is also possible [143]. True over-fitting in learner selection occurs when considering new learners for selection actually reduces the true performance of the final learner selection.

4.7 Relationship with publication bias in other fields

Selection bias is a consequence of the broader phenomenon of publication bias, the name given to what occurs when published research is systematically unrepresentative of a population of ideal completed studies. In 'classical' publication bias, researchers looking for group differences or associations conduct generate many feasible results but then only report those that are favourable [147].

The various results may be generated by independent groups of researchers using independent datasets, or they may be produced by a single group using different variations of experimental or analytical parameters. While none of these parameters settings may be invalid in themselves, the selection from multiple results provides a greater likelihood that an impressive

result can be produced on a particular dataset.

When a single hypothesis is considered, publication effectively increases the false positive rates of statistical tests. In a field where multiple hypotheses are tested, publication bias increases the fraction of published findings that are false. This is particularly the case when a large number of hypotheses are tested without strong prior justification, as this makes the prior probability of a non-zero effect low [41].

In ANA studies, researchers may not only adjust their samples and methodology but also the specification of the learner. This may be viewed as either testing additional hypotheses in search of a favourable result (learner 1 is better, or learner 2 is better, or...) or varying an experimental parameter. While publication bias causes false positives and effect size inflation in group difference studies [12], in ANA studies it causes biased assessment of learner performance.

In his influential paper of 2005, "Why Most Published Research Findings Are False", Ioannidis identifies key underlying risk factors that increase the probability that publication bias will play a key role in a field [41]. These include the following:

1. small sample or effect sizes (corresponding to high noise relative to measured effects);
2. great flexibility in designs, definitions, outcomes, and analytical modes;
3. great number and lesser pre-selection of tested relationships;
4. great interest on the part of the researchers in obtaining a particular result; and
5. many research groups chasing an objective (such as statistical significance).

The first four of these are very common across all ANA research. As demonstrated by the long length of my supplementary document, the last is certainly also present in AD classification research.

Summary

I have provided a detailed review of selection bias in AD ANA performance results, and described how it is distinct from other sources of bias. Selection bias is caused by the selective reporting of performance results based on their relative values. This can happen in the work of single researcher or group who only submit the most promising results for publication, or in the work of an entire research community working on single shared dataset such as that of ADNI. Selection bias is both a form of publication bias and regression to the mean.

Chapter 5

An empirical investigation into selection bias in AD classification

In this chapter, I shall describe an empirical investigation I have conducted into selection bias in AD classification. The chapter is organised as follows: the motivation for the investigation are discussed in section 5.1, the materials and methods (including the experiment design) are described in section 5.2, the results are presented in section 5.3, and the implications for AD classification research, and ANA in general, are discussed in section 5.4.

The experimental design used in this investigation is based on that used in a preliminary investigation presented at the international conference on medical image computing and computer assisted intervention (MICCAI) in 2014 [35].

This study presented in this chapter has been published in the journal NeuroImage:Clinical [149].

5.1 Motivation

As discussed in section 4.1, bias in performance estimation should be an important concern for ANA researchers, as biased results can lead to the introduction of inferior clinical decision systems. Bias due to changes in training set size is relatively predictable, and the fact that it is typically pessimistic rather than optimistic makes it a more minor consideration. Bias due to population shift is likely to remain a challenge, though several studies have explored the problem of generalisation between different populations [22, 43, 142]. While various studies have warned about the potential for bias in the selection of algorithm parameters in a single study [23, 145], the problem of selection bias due to the selection of learner *specifications*, either by a single group or an entire field, has received relatively little attention.

AD classification is probably the most well studied learning problem in ANA [27, 29]. As demonstrated by my supplementary document, a vast number of learners have been developed

and evaluated in the search for improved performance [7]. This should make the issue of selection bias particularly concerning for AD classification researchers, as their field is one of those most likely to suffer from it to a significant degree.

The study of this chapter aims to provide plausible estimates for the degree of selection bias that may be present in the AD classification literature, and to better understand the relationship between selection bias and its key determining factors, including

- sample size,
- classification task,
- CV strategy, and
- the number of learner specifications considered for selection.

By doing so, it should be possible to make useful recommendations to improve validation practice in future AD classification research and to aid in the critical assessment of the performance results that are already published.

The study uses the dataset of the Alzheimer's disease neuroimaging initiative (ADNI), which is the same one used in the majority of AD ANA studies [24]. It uses a variety of simplified learners based on real pipeline components and a resampling based experiment design to provide what may be viewed as a loose simulation of the selection processes occurring in the actual research field.

5.2 Materials and methods

The organisation of the materials and methods is as follows: in section 5.2.1, I describe the subjects and images I used; I then describe how I built the learners in section 5.2.2 and the design of my experiments in sections 5.2.3 and 5.2.4.

5.2.1 Subjects and Imaging Data

Imaging and clinical data were obtained from the ADNI database, details of which are provided in section 2.3.

All subjects were designated as healthy control (HC), AD or MCI at the time of the baseline scan, and were subsequently reassessed at time-points during follow-up. Inclusion criteria for HC subjects are mini mental state examination (MMSE) scores between 24 and 30, a clinical dementia rating (CDR) of 0, non-depressed and non-demented. Ages of the HC subjects were roughly matched to those of the AD and MCI subjects. For MCI subjects, the criteria are an MMSE score between 24 and 30, a memory complaint, objective memory loss measured

by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. AD subjects were identified by an MMSE score between 20 and 26, CDR of 0.5 or 1.0, and the NINCDS/ADRDA criteria for probable AD [150]. For the purposes of this study, MCI subjects were considered stable if they had an assessment up to or beyond the subsequent 24 months follow-up period in which they were not given a diagnosis of AD. MCI subjects were considered progressive if they were given a diagnosis of AD at any point during the follow-up period. MCI subjects whose progression status could not be determined were excluded from the sample used for experiments. Subjects with a suspected dementia aetiologies other than AD were excluded from my classification experiments.

In this study, I used T1-weighted structural MR images from the baseline time-point alone. This is the most commonly encountered imaging setting in AD classification research [24], in part because it is one of the most easily achievable in clinical practice. It also offers a larger total sample than any other imaging choice, something that is crucial when using subsampling designs such as the one in this experiment. Because the design limits one to using no more than half the available full sample in a given CV experiment, it is crucial that the full sample is as large as possible to ensure that the subsamples used for CV are of a realistic size. Larger samples should also provide more accurate estimation of performance quantities and bias.

In order to further enlarge the sample, I decided to include images from both 1.5 and 3.0 Tesla scanners. This is not unprecedented, and it may be a sensible option for future diagnostic tools in clinical practice where the losses in performance due to heterogeneity may be outweighed by the gains due to increased training sample sizes [43]. All images were post-processed to correct for gradient warping, B1 non-uniformity and intensity non-uniformity and underwent phantom-based scaling correction. I conducted my own quality control assessment in addition to that provided by ADNI to cover subjects for whom no quality assessed images were available. Where back-to-back images were available for the baseline time-point, the one with the superior quality score was selected.

The selection criteria described yielded a total of 1437 subjects to be used in the image processing. Among these, there were 372 HC subjects, 252 AD subjects, 230 stable MCI subjects (MCIs) and 135 progressive MCI subjects (MCIp).

5.2.1.1 Image processing

A sample specific groupwise space was created using iterative affine and then B-spline registration using the publicly available NiftyReg package[1]. Tissue segmentation and atlas propagation algorithms (more details in sections 5.2.2.2 and 5.2.2.3 respectively) were applied to all images in their native space. Tissue segmentations and atlas labels were propagated to the groupwise space, where the latter were combined to produce a sample specific group atlas.

5.2.2 Learners

Due to the constraints of implementation and computational time, it would not have been possible for me to include many of the highest performing classification learners from the literature. Instead, I tried to produce a large but plausible set of learners based on pairings of one of the 48 feature sets described in 5.2.2.1 and one of the 6 classification algorithms described in 5.2.2.6. Because one of the algorithms (random forest) cannot be combined with 24 of the (kernel-based) feature sets, there are 264 learners in total.

5.2.2.1 Feature sets

I produced my 48 feature sets in the same combinatorial way that I produced my learners. Each feature set is a combination of some imaging descriptor (see section 5.2.2.2), one of two atlases used to interpret that descriptor (section 5.2.2.3), and some way of using the atlas to perform a knowledge-based feature selection (section 5.2.2.4). Note that not all combinations of options were possible.

5.2.2.2 Imaging descriptors

All imaging descriptors were produced using no more than one of two grey matter (GM) tissue concentrations and one of two atlas segmentations. GM concentration maps are one of the most fundamental tools for the study of structural changes in the brain; they have a key role in the voxel-based morphometry that has become the established tool for group difference studies in structural neuroimaging. They were the first image descriptors considered for AD classification [34], and they are still studied frequently [25, 78]. Where I did not use GM concentrations, I used the volumes of atlas regions. Though these have not been used as commonly as GM measures, they do provide a straightforward description of brain atrophy that can be used to classify neurological diseases associated with it [151].

The imaging descriptor can be divided into two groups: *primal* descriptors representing the values of a quantity in each region of an atlas in the native space, and dual or *kernel* descriptors represented by kernel matrices computed from voxelwise intensity scores in the groupwise

[1] `http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg`

space.

Primal descriptors Regional GM loads were calculated by summing tissue concentrations over regions of the atlas in the *native space* multiplied by the volume of the voxels. These loads should reflect both the concentration of GM and its volume.

1. ***SPM GM loads***. I used the publicly available SPM12 package[2] to provide tissue concentrations maps.

2. ***GIF GM loads***. The geodesic information flow (GIF) algorithm [152] used to produce the Neuromorph parcellations (see section 5.2.2.3) jointly estimates various tissue concentrations maps. I included the resulting GM maps as an alternative to those of SPM.

3. ***Atlas region volumes***. The volumes of atlas regions were normalised by total intra-cranial volume as measured by the union of relevant SPM tissue maps.

Kernel descriptors All kernel descriptors were produced from one of the same two GM maps used for the primal descriptors after they had been mapped to the *groupwise space*. Kernel matrices were computed separately for each region of the group atlas, and then later combined by summation. There were three levels of further processing possible in the groupwise space, producing to a total of six kernel descriptors.

1. ***No further processing*** as the simplest option.

2. ***Modulation*** by the Jacobian determinant of transform from the native to the groupwise space.

3. ***Smoothing*** (performed in addition to modulation) using an isotropic Gaussian kernel of 2.0 mm standard deviation (4.7 mm full width half maximum). This aims to compensate for registration errors and the spatial variation of atrophy patterns. The choice of 2.0 mm was intended to be a middle-of-the-road choice.

5.2.2.3 Atlases

Through its interaction with the choice of imaging feature type, the choice of atlas can be an important determinant of a learner's performance [78]. I included two choices of brain atlas in my feature sets. These offer competing definitions of what precisely constitutes an anatomical brain region, and determine which areas of the brain will be used in predictive modelling.

[2] `http://www.fil.ion.ucl.ac.uk/spm/doc/`

Hammers. This is an atlas of 83 regions described in [99, 100], with 30 manually labelled reference images available online[3]. The labels of these images were propagated to the space of my images and fused using the STEPS algorithm [153]. Notably, many of its regions comprise both cortical grey matter and the white matter beneath it.

Neuromorph. This is an atlas of up to 141 brain regions provided by the commercial company Neuromorphometrics, Inc. under academic subscription. Notably, in this atlas, cerebral white matter is held in separate compartments from cortical grey matter. I use 35 labelled reference images originating from the OASIS project that were made available for the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labelling[4]. These were propagated and fused using the GIF algorithm described in [152].

A group definition for both atlas types was produced by propagating the native space atlas labels from all 1437 images and combining them by majority vote.

5.2.2.4 Spatial restriction/feature selection

The selection of relevant features can be used to remove variability in the data unrelated to the class discrimination problem. Knowledge-driven feature selection techniques have been shown to be superior to data-driven ones [105], and they have the added advantage of being computationally inexpensive. I consider two types of spatial restriction that implement a knowledge-driven feature selection.

A symmetry constraint enforced by combining, as an average, each pair of features related to a brain region occurring in each hemisphere. While the atrophy associated with AD may not be symmetric [154], the modes of atrophy that are most informative for classification may be. Due to the difficulty establishing a voxel-to-voxel correspondence between the hemispheres, the constraint was not applied to kernel-based feature sets.

An exclusive focus on the temporal lobes justified by their well established role in AD [155, 156].

Zero, one, or both of these were applied to produce a feature set.

5.2.2.5 Standardisation

All kernels were scaled so that the median inter-point distance in the whole sample was one. While this linear scaling should have no impact on performance, it is mentioned because of its interaction with the `C` parameter in the SVM algorithm (see section 5.2.2.6). All primal

[3] `http://brain-development.org/brain-atlases/`
[4] `https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Main_Page`

features were standardised by subtraction and division to ensure a zero mean and a unit variance. This standardisation was performance using the full sample, rather than seperately for each training set. As discussed in appendix A, as with the groupwise registration used here there is some potential for the introduction of a positive bias. However, because this transformation is relatively simple and the number of items is relatively large, the bias introduced is unlikely to be significant.

5.2.2.6 Classifier algorithms

Unless otherwise stated, all classifier algorithms were implemented in `C++` by me. The two more simple algorithms were included to ensure a diverse collection and to illustrate to what extent the choice of algorithm is important in the construction of learners. A more detailed description of the relevant learners can be found in section 2.6. A summary of the relevant implementation details is presented below.

SVM. The support vector machine. I used the publicly available `libsvm` package [157] with a linear (precomputed) kernel, as is commonly preferred for the high dimensional classification problems of neuroimaging [32]. The `C` parameter was selected from the values $2^{-2}, 2^{-1}, \cdots, 2^{4}$ using nested two-fold CV with five repeats. Results in preliminary experiments were essentially identical if the range of `C` values considered was expanded at either end of this range.

RF. A random forest (RF) classification algorithm based on the original specification by Breiman [123] with the parameter `mtry` set to the rounded square root of the number of features. As I know of no obvious extension of RF to make use of kernel descriptions, these combinations are omitted from the set of learners considered. I used 100 trees; increasing this number in preliminary experiments produced essentially identical results.

LDA1. Linear discriminant analysis classification. In LDA1, I used the standard formulation with the standard maximum likelihood covariance estimate for the distributions of the two classes. For primal feature sets only, the Ledoit-Wolf lemma [116] was used to provide a shrinkage estimator of the covariance (this is relatively standard, and appears in the sklearn learning package[5]). The threshold used for classification was based on a Gaussian model using class prior probabilities derived from the training set.

LDA2. An alternative version of LDA where the full sample covariance matrix was used, rather than the sum of the covariance contributions from each of the subject groups. This pro-

[5] http://scikit-learn.org/0.16/modules/generated/sklearn.lda.LDA.html

duces a biased, but slightly more stable estimator of the covariance. I have found this to perform better than LDA1 in synthetic high dimensional problems.

NC. The nearest centroid (NC) algorithm. This can be seen as something of a 'control' for the more complex linear methods (LDA,SVM) Comparison with these will show how important the estimation of the covariance structure is.

KNN. The K-nearest neighbours (KNN) algorithm. Because the distances between points may be obtained straightforwardly from the kernel matrix, KNN may be applied to kernel-based problems. The number of neighbours was selected from the set of odd numbers less than one third of the size of the training set using nested two-fold CV with five repeats.

The C4.5 decision tree algorithm (see section 2.6.2) was originally included in this list, but was excluded after experiments revealed it to rarely perform better than chance.

5.2.3 Cross validation and performance measures

I considered stratified repeated K-fold cross validation (RKCV) as an estimator of learner accuracy, this is the most common strategy in AD classification research. The performance estimate of a single repeat of K-fold cross validation (KCV) is the fraction of subjects that were classified correctly, and the final estimate is averaged over one or more repeats entailing a randomly produced partition of the data. The numbers of items that appear in each fold are determined exactly as in the publicly available libsvm package [157]. That is, where there are l_c items in a sample of class c, the kth fold contains $\lfloor l_c/K \rfloor + \lfloor (l_c \bmod K)/k \rfloor$. From here on, I shall use ExKcv to denote EKCV using E repeats of K-fold CV.

5.2.4 Subsampling experiment design

In order to estimate the bias associated with selection of learners based on performance, one needs to obtain both biased and unbiased performance estimates. To do this, I created the subsampling design described in figure 5.1 and defined below.

1. Two disjoint subsamples, respectively called the left and right, of a specified sample size and class composition are drawn randomly and without replacement from the full set of available samples.

2. Some form of (repeated) KCV is applied to estimate the performance of the learners in the left and right subsamples separately. All learners are compared in parallel alongside one another.

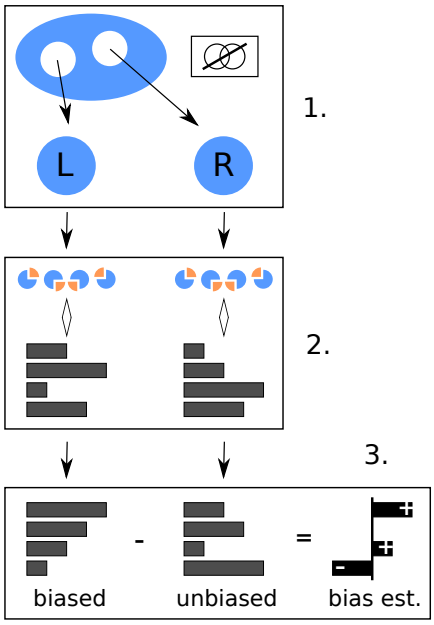

Figure 5.1: One iteration of the full experiment design described in section 5.2.4. In step 1, two disjoint subsets labelled left and right are drawn from the full available dataset. In step 2, some form of CV is used to produce two independent performance estimates for a number of learners. In step 3, these are ranked by their values in the left subset. The difference between the two estimate sets is then taken as an estimator of the bias associated with the different rank positions.

3. The learners are ranked by their performance in the left dataset. The *n*th ranked performance measurement in the left dataset (henceforth, the in-sample estimate) is a biased estimator for the performance of the learner that obtained that rank (see section 4.3), but the corresponding measurement in the right dataset (henceforth, the out-of-sample estimate) is not. The difference between the two is an (unbiased) estimator for the selection bias associated with the *n*th rank position.

4. The last step is repeated, but this time the roles of the left and right datasets are reversed. The average of the two resulting bias estimates is taken.

This process is repeated 2000 times using different random partitions into left and right subsets, and the bias estimates are averaged together to provide greater stability. The number 2000 was chosen to ensure that the choice of random partitions had no appreciable effect on the outcome of the experiment in preliminary work.

I used this design to investigate two classification tasks: the discrimination of subjects with AD from healthy controls (henceforth, AD detection), and the discrimination of MCI subjects who went on to progress to AD in a 24 month interval from those who did not (henceforth, MCI prognosis). Both tasks are conducted in samples containing only the two subject classes to be discriminated.

In order to investigate the effects of sample size and CV strategy, I repeat this experiment procedure many times while varying these parameters. When varying the sample size used for the left/right subsamples, I kept the ratio of positive and negative classes fixed at 2 : 3, a ratio which closely approximated that in the full available sample for both tasks (see section 5.2.1). For AD detection, a subsample of size 50 had 20 AD subjects and 30 controls. Similarly, for MCI prognosis, a subsample of size 50 comprised 20 progressive subjects and 30 who remained stable. Under this class balance constraint, sample sizes were varied from 30 up to the maximum permitted by the full available sample in steps of 10. I repeated all experiments using ExKcv strategies with $K \in \{1,2,...,24\}$ and $E \in \{1,2,...,24/K\}$, producing bias estimates for both tasks with a variety of strategies.

5.2.4.1 Number of learners considered

I also investigated the effect of the number of learners considered on selection bias. The chance of the learner obtaining the rank p out of n (with $p = 1$ denoting the lowest performance) being ranked the highest out of k_{select} learners selected randomly without replacement is

$$P_{\text{highest of } k_{\text{select}}} = \frac{\binom{n-p}{k_{\text{select}}-1}}{\binom{n}{k_{\text{select}}}}. \tag{5.1}$$

By using this formula to combine the biases associated with all n ranks, I was able to measure the average bias associated with the best performing learner out of all $\binom{n}{k}$ possible learner subsets of size k.

5.2.4.2 Decision power

Selection bias is intimately related to an experiment's ability to correctly identify the best of several learners. To characterise this, I considered the rate at which a CV experiment was able to identify the superior of two learners based on their measured performance. Specifically, I considered the 'posterior probability' that one learner performance was truly better than another based on an observed performance difference in a given experiment. I call this probability the 'decision power'.

To estimate this, I took the average performance of a learner across all 2000 experiment pairs as a gold standard for its true performance. I then computed the fraction of times that a pairwise performance difference in an individual experiment had the same sign as the corresponding gold standard pairwise performance difference. By including only those observed performance differences that had a magnitude within a certain interval, I could compute the posterior probability that the ranking of the two learners was correctly estimated in an experiment after observing a performance difference in that interval.

5.3 Results

Where the curves of multiple sample sizes appear in the plots of this section, these have been selected to display a representative range of behaviours. Where a single CV strategy is presented, this is 4x6cv. This was chosen for its relatively low bias and its intermediate fold number as compared to other strategies considered. Where both in-sample and out-of-sample accuracies are presented, the former is biased, while the latter can be regarded as accurate.

5.3.1 Accuracy of learners

Here, I present the average accuracy of the learners in the largest of the samples considered in the subsampling design in figures 5.3 and 5.4. A key to understanding these is presented in figure 5.2. Note that the 3:2 ratio of negative (HC/MCIs) to positive (AD/MCIp) classes should give a "null" accuracy of 60% for both tasks.

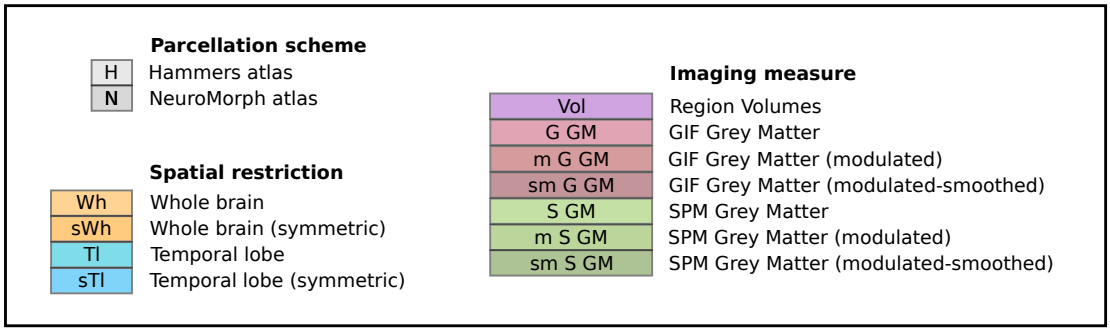

Figure 5.2: Key to accuracy figures 5.3 and 5.4. Each feature set is described by one label from each group.

My design allows me to produce an unbiased variance estimate for the performance as estimated in a single CV experiment [35]. This in turn allows me to produce a conservative (upwardly biased) estimate for the variance of my final performance estimate (by falsely assuming the average of all 2000 CV experiment pairs has the same variance as the average of a single pair). This conservative estimate gives standard deviations in the range 1 to 3.5% for all accuracy estimates displayed in this section. For the AD detection task (figure 5.3) most of these are under 2%; for MCI prognosis (figure 5.4), most of these are above 2%.

Though I present them primarily to demonstrate the plausibility of my methods, I can also make some comments on the accuracies observed. These span a wide range in both tasks, with those in AD detection spanning the range 70-90%, and those in MCI prognosis spanning the range 60-70%. This is towards the lower end of the spectrum of published results [14, 24, 88].

A loose hierarchy of classifier algorithms is evident in the AD detection task, with SVM, LDA and RF consistently performing better than NC and KNN with a given feature set. This is not so much the case for MCI prognosis, where the ranking of algorithms is highly dependent

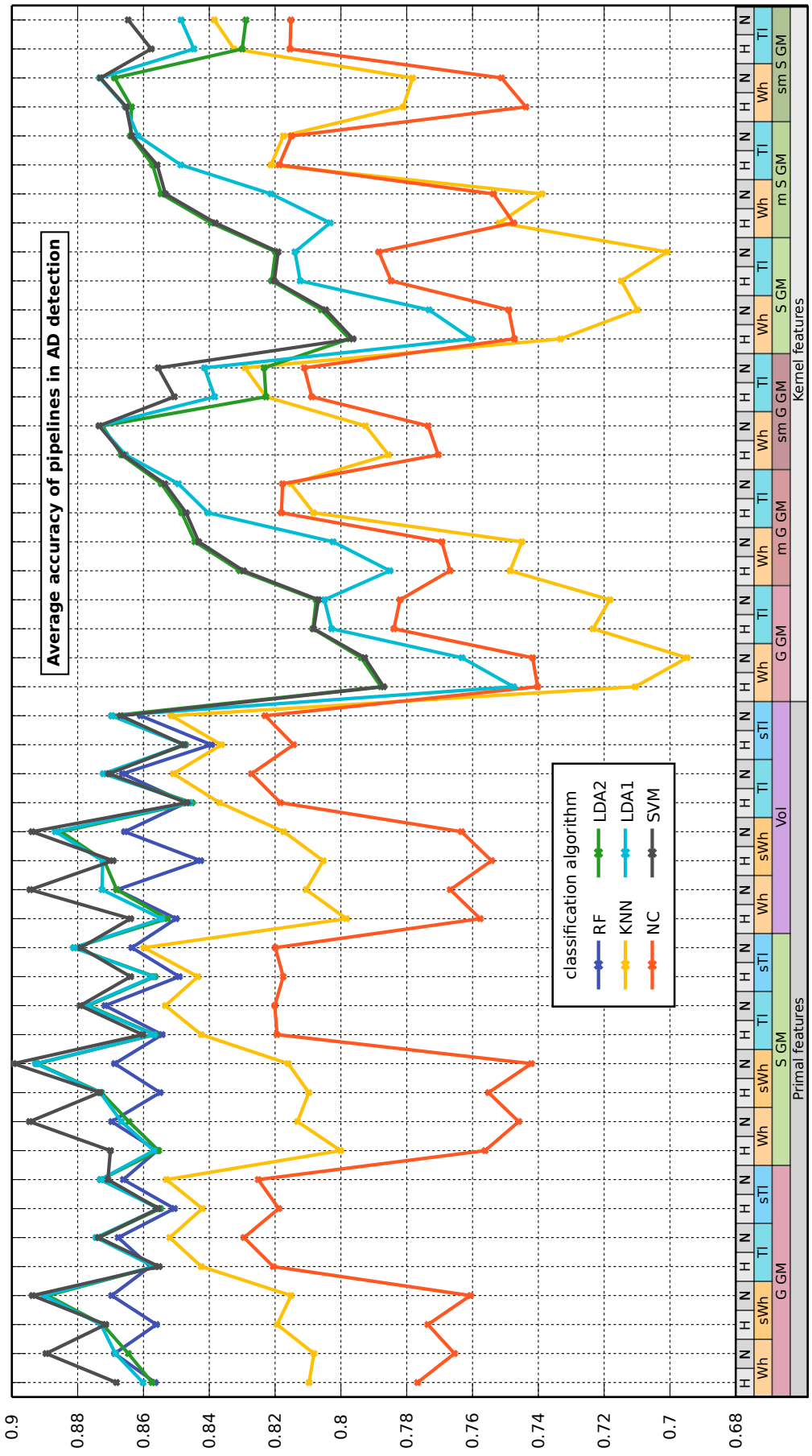

Figure 5.3: Average accuracies of AD diagnosis learners measured using 4x6cv and 300 subjects. See key in figure 5.2.

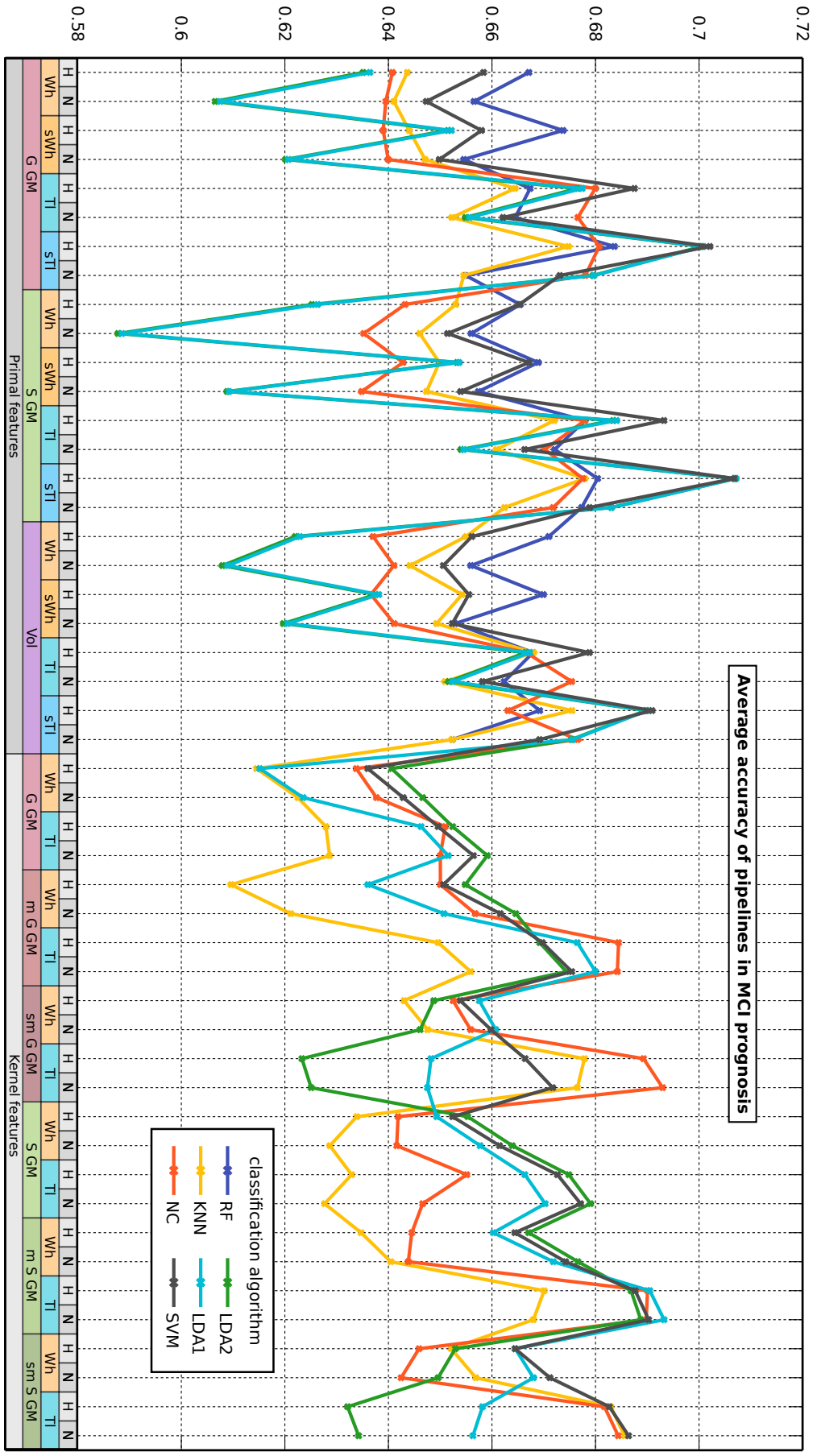

Figure 5.4: Average accuracies of MCI prognosis learners measured using 4x6cv with 160 subjects. See key in figure 5.2.

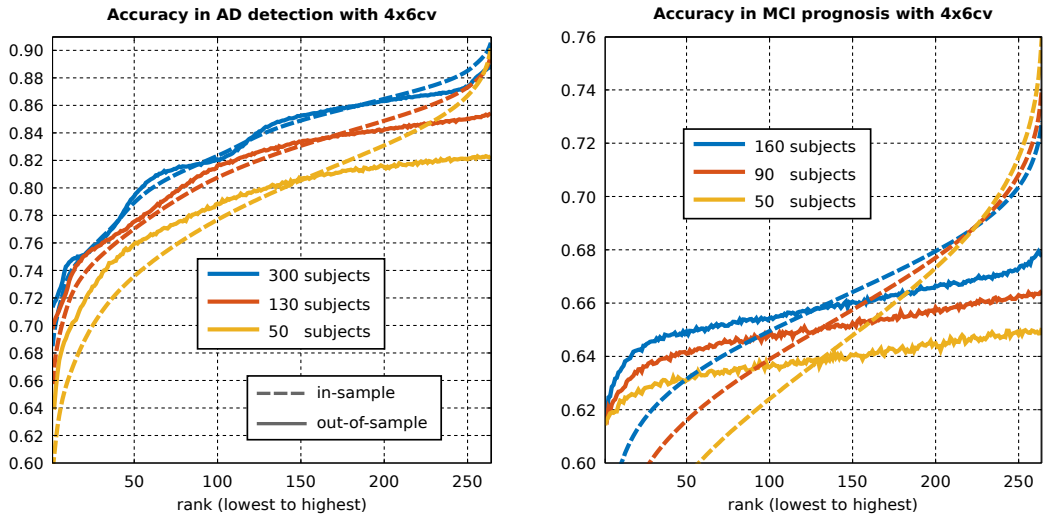

Figure 5.5: Illustration of the effect of ranking on performance estimation. In-sample and out-of-sample performance estimates for both tasks are presented above, with the bias (the difference between the two) visible below.

on the feature set. This may be due to the more difficult nature of the task confounding the fitting of more complex models. In both cases, clear repeating patterns are visible, with GIF GM, SPM GM and region volumes producing similar results with the same regional restrictions.

5.3.2 Bias as function of rank

In figure 5.5, I present my estimates for the selection bias associated with different rank positions using 4x6cv. The in-sample performance estimates presented in the upper portion of the figure are those used to rank the learners, while the out-of-sample estimates are those from CV in the independent disjoint sample. The biases, presented in the lower section of the figure, are the differences between the two.

The selection bias has a loosely sigmoid shape, with low ranking learners in-sample accuracy estimates having a negative bias, and high ranking learners estimates have a positive bias. The bias curve for AD deviates from this shape at larger sample sizes, which may be due to the greater stability of learner rankings. At all ranks, the magnitude of the selection bias is lower at greater sample sizes. In MCI prognosis, most of the difference between the lowest and highest ranking learners' in-sample performance estimates is due to selection bias. For example, when using 160 subjects, the in-sample difference between the highest and lowest performing learners was on average roughly 16%, the out-of-sample difference between the two was on average only roughly 6%. For AD detection, this is not the case, with the majority of the difference between lowest and highest ranking learners being repeatable in an independent sample.

5.3.3 Bias as function of CV strategy and sample size

This section describes the observed relationship between the sample size and strategy used for CV and the bias associated with the best performing learner. Figure 5.6 illustrates the relationship between bias, choice of K in KCV, and the parameter E determining number of full KCV repetitions. The upper part of the figure demonstrates the beneficial effect of additional KCV repeats, which reduce bias by reducing the variance of performance measurement. The choice of K was made to illustrate this effect both for computational convenience and because the gains associated with further repeats are more dramatic when using smaller K values. The lower part of the figure illustrates the effect of K on bias. In order to consider this effect separately from the benefit offered by additional train-test cycles, E is adjusted simultaneously to ensure that all strategies considered entail a roughly equal amount of computational effort (as in [158]). This makes the comparison of strategies more practically relevant to an experimenter considering the right way to make use of limited computational resources[6]. In all cases, selection bias is roughly proportional to the inverse square root of the sample size. The axes have been rescaled to illustrate this relationship.

The relationship between sample size, expected in-sample performance, and true out of sample performance of the highest performing learner (the gap between these two being the bias) is illustrated in figure 5.7. As expected, out-of-sample performance increases with sample size because there are more items available for training. However, *the decrease in bias associated with larger samples is large enough to cause in-sample performanc estimates to actually decrease as sample size increases*, an effect that would seem paradoxical in the absence of the selection process. This effect is so strong that the highest expected in-sample performance estimates are observed at the smallest sample sizes in both classification tasks. Indeed, it is only in AD detection that increasing sample size can ever lead to increasing in-sample performance estimates for the best performing learner. This is because it is only in AD detection that genuine improvement in performance (as measured out-of-sample) with sample size can be great enough to outweigh the associated reduction in bias.

Also notable in figure 5.7 is the effect of the number of KCV repetitions E on expected in-sample and out-of-sample performance. Higher repetition numbers reduce bias, and so decrease apparent in-sample performance at all sample sizes. Though the effect is small, the additional repeats also provide a more robust selection of truly better learners, leading to a small increase in the out-of-sample performance estimates.

[6] Just as it is unfair to compare 1x5cv to 10x5cv, it is unfair to compare 1x5cv with 1x50cv.

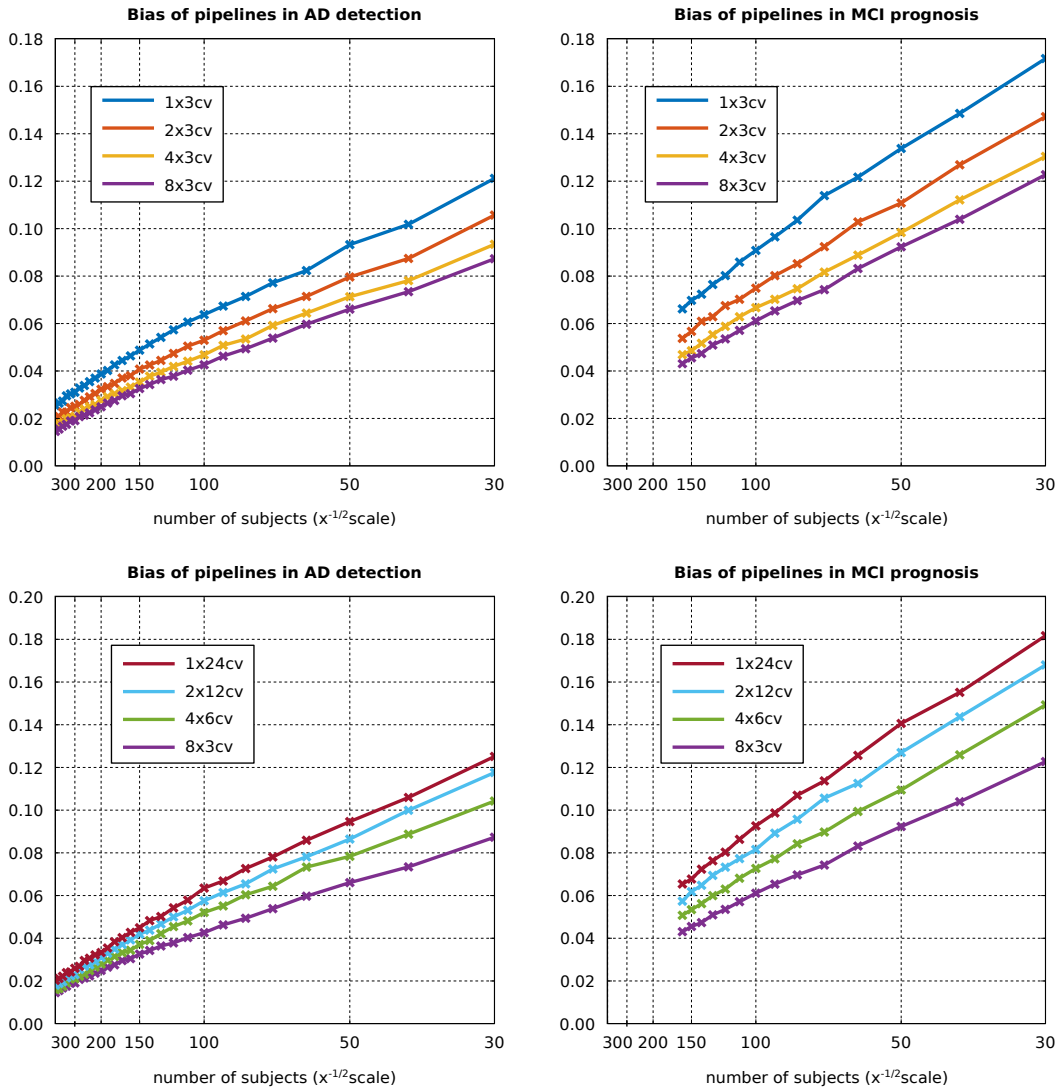

Figure 5.6: Effect of sample size and KCV strategy on the bias associated with the highest performing learner.

5.3.4 Bias as function of the number of learners considered

Figure 5.8 illustrates the effect of k_{select}, the number of learners considered, on the bias associated with selecting the best. When only one learner is considered, both in-sample and out-of-sample performances are the same, as there is no selection bias. As the number of learners considered increases, one can see the out-of-sample performance increase as the degree of opportunity for real improvement grows. The in-sample performance of the best performing learner grows at a much faster rate, reflecting the faster growth of the selection bias. The degree of this divergence is dependent on the sample size and the classification task. For MCI prognosis, it can be seen that most of the improvement in the accuracy of the best learner seen after increasing the number of learners considered is due to bias alone. For AD detection, this is only true at smaller sample sizes.

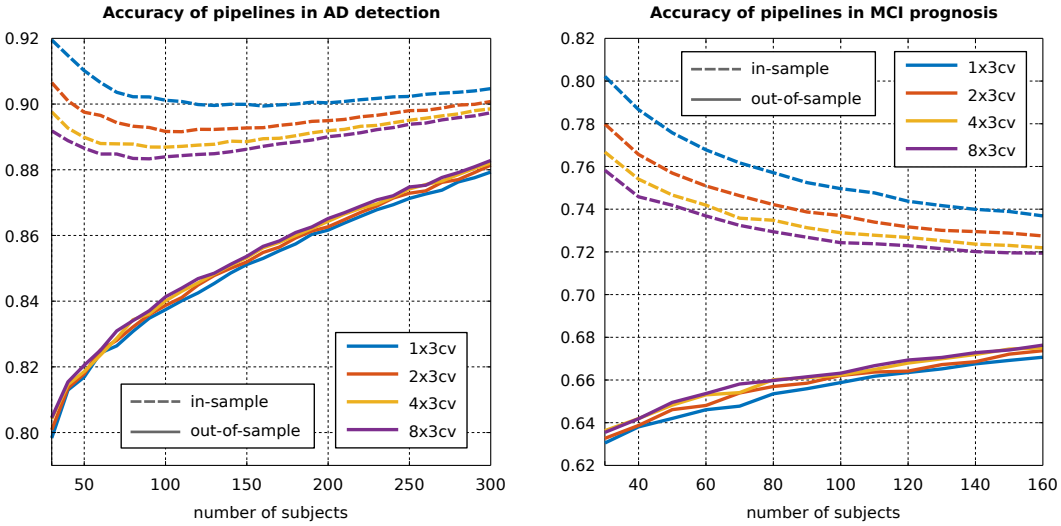

Figure 5.7: Effect of sample size and CV strategy on the performance of the learner with the highest observed performance. The upper curves represent the in-sample performance measures available to an experimenter, while the lower curves represent the true out-of-sample performance. Note how, from the viewpoint of an experiment, increasing sample sizes may lead to an apparent drop in performance.

For both in-sample and out-of-sample accuracies, there is a quickly diminishing return in performance as more learners are considered. The bias (the difference between them) similarly increases slower than the logarithm of the number of learners. This is consistent with the results for the simple model described in 4.4.1.

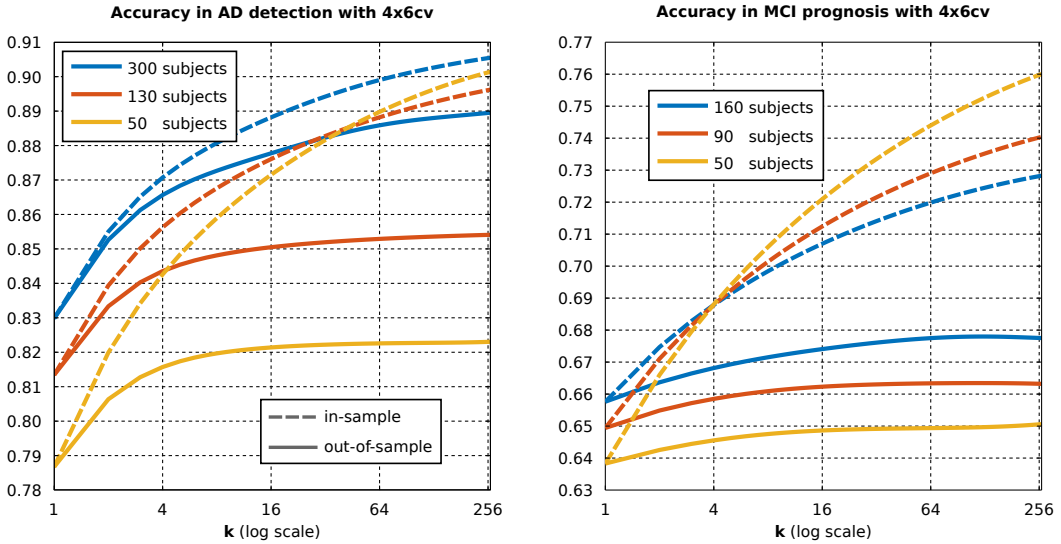

Figure 5.8: Effect of number of learners considered on the (biased) in-sample accuracy and the (unbiased) out-of-sample accuracy associated with the highest ranking learner. The in-sample estimate can be seen as representing the apparent progress associated with learner optimisation, while the out-of-sample estimates represents the true progress. It can be seen that much of apparent progress associated with an increasing number of learner options is spurious (i.e. due to increasing bias alone).

5.3.5 Bias as a function of precision in performance estimation

The use of disjoint subsets in the experiment design allows me to measure the variance of quantities without bias [35]. For all experimental settings (sample size, CV strategy), I used all experiment repetitions to measure the variance of the difference in performance between each pair of learners. I took the average over all pairs to produce an indicative expected variance in a pairwise performance difference. I then took the square root of this quantity as an indicative measure of the 'noise' in performance ranking.

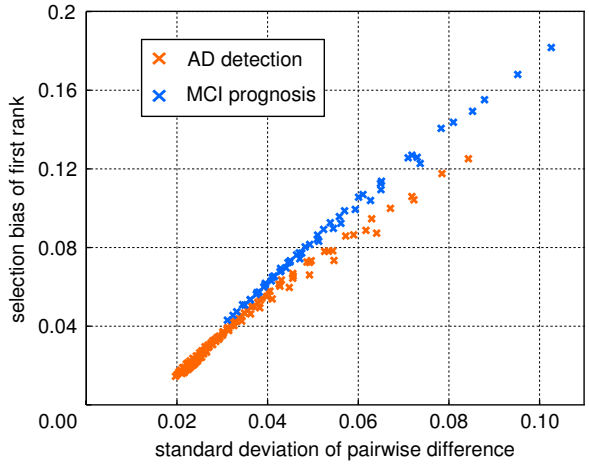

Figure 5.9: Illustration of the relationship between the precision in measuring the differences in performance and selection bias. Each point represents a given setting of sample size and number of KCV repetitions. The standard deviation represented by the X-axis is the square root of the representative variance of pairwise performance differences describe in section 5.3.5.

The relationship between the indicative noise and the selection bias associated with the highest ranking learner is illustrated in figure 5.9. Although the representative noise measure does not fully explain the level of bias, the relationship is approximately linear. This emphasises the importance of low variance in performance estimation for the reduction of selection bias.

5.3.6 Decision power

Figure 5.10 demonstrates the reliability in the case of the two learners, and shows its relationship with the observed difference in performance. It is clear that greater sample sizes improve decision power. The 50 subject examples in both tasks show that AD detection learner rankings at a given sample size are more reliable than those for MCI prognosis learners at the same sample size. Note that, for all observed performance difference bands, the decision power in AD detection with 110 subjects is greater than that in MCI prognosis with 160 subjects. The magnitude of performance difference required to ensure a 90% correct ranking surprised me; for MCI prognosis experiments with 160 subjects, an average performance difference of 7% is required to reduce the chance of an incorrect ranking to 5%. With 90 subjects, this jumps up to

14%. With 50, differences sufficiently large did not occur a great number of times.

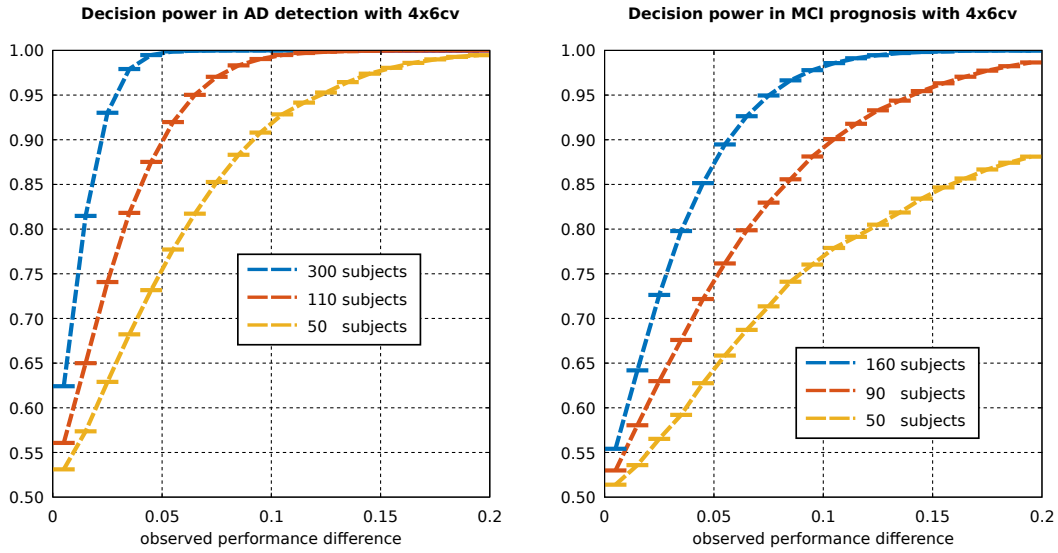

Figure 5.10: Illustration of the decision power associated with 4x6cv in both classification tasks. This is the probability that learner B is truly superior to learner A, given that B performed X better than A in a CV experiment, where X is the value of the X-axis. Horizontal bars represent calculation intervals. For $i \in \{1,2,...,20\}$, interval i contains difference d when $(i-1) < 100|d| \leq i$.

5.4 Discussion

The investigation of this chapter has demonstrated that selection bias should be present in AD classification performance results and that the magnitude of this bias should be comparable to the observed improvement in performance associated with learner specification optimisation. As anticipated by the simple model of section 4.4.1, the bias increases with level of 'measurement noise' and the number of learner specifications considered (see section 5.3.5). It appears to be approximately proportional to the inverse square of the sample (see figure 5.6), which itself should be approximately proportional to the standard deviation of the measurement of learner performance. The CV strategy used also plays a role in determining bias, with higher values of E in ExKcv reducing measurement noise and bias. Finally, bias is also dependent on the classification task considered, being higher in MCI prognosis than in AD diagnosis when all other factors are identical (figures 5.6- 5.7). This could be explained by a simple binomial accuracy measure model, as the lower accuracies (such as that in MCI prognosis) should produce a higher variance in the mean accuracy measured in a test set.

Crucially, selection bias is responsible for a large fraction of the apparent improvement associated with selection from an expanding pool of learner specifications, even at sample sizes that can be considered relatively large (figure 5.8). This is particularly striking in the case of MCI prognosis; even with 160 subjects, the largest size considered here, bias is responsible

for more than two thirds of the in-sample performance improvement when moving from a pool containing only one candidate learner to a pool containing 264 (in-sample performance rises from 66 to 74%, but out-of-sample performance only rises from 66 to 68%).

Another interesting result is the high difference in observed performance required to ensure that two learners are correctly ranked. As discussed in chapter 9, ANA learners are often compared on the basis of point estimated performance alone. Many studies using similar sample sizes and CV strategies reach firm conclusions on the relative merits of learners based on smaller differences in point estimated performance than those that would be necessary to secure a high posterior probability of a correct ranking in this experiment. If my results are representative of the field, the conclusions of these studies may be less certain than they are held to be.

While this study is limited to AD classification, it is reasonable to expect that selection bias should behave in a qualitatively similar way in any other applied field of machine learning where a large number of researchers use CV to search for superior learner specifications using limited collections of items.

5.4.1 Risk factors for bias

My results suggest a need for caution in the interpretation of published results, particularly in contexts containing one or more of the following risk factors:

- the classification task is MCI prognosis (or some other difficult task);
- the sample used to demonstrate CV performance results is small;
- multiple learner specifications are evaluated in a study, and results are reported as "up to X"; or
- a high performing learner involves many steps with adjustable settings, and it is not clear how those settings have been chosen.

This list of factors may be viewed as a specific reinterpretation of the more general factors identified in [41].

A corollary of this consideration is that some of the apparent performance advantage associated with multi-modal learners over single modality alternatives [28, 46] may be due to selection bias, as multi-modal learners entail a greater degree of complexity and are necessarily assessed using smaller validation samples.

5.4.2 Evidence of bias in the AD classification literature

The nature of selective reporting makes it difficult to assess the level of bias in the literature directly. One exception to this is found in challenges such as CADdementia [25], which report both biased in-sample performance estimates and unbiased out-of-sample performance estimates. As can be seen in figure 7 of the relevant paper [25], all 29 contestants overestimated the accuracy of their submitted AD classification predictors. Though it was not based on an AD classification problem, a similar pattern is apparent in the MICCAI 2014 machine learning challenge[7], where all but two of the 48 submissions overestimated their performance. Learner developers may be less concerned about providing biased performance estimates when participating in a challenge than when conducting a standalone study. As such, it should be noted that selection bias will be larger in challenges than in the literature in general.

Another way selection bias may make itself apparent is through the relationship between sample size and reported performance. As can be seen in figure 5.7, despite the expected positive association between training set sizes and learner performance [35, 105], in the presence of a selection process, smaller sample sizes can actually facilitate more impressive (apparent) performance results. This is analogous to a phenomenon seen in group difference studies [12], where small sample sizes provide more scope for large in-sample effect sizes to emerge by chance.

In order to provide a representative sample of studies, I selected from those used in the recent review of Arbabshirani et al. [29]. From that review, I selected all those studies that consider AD detection or MCI prognosis tasks using sMRI data alone. This restriction was intended to avoid a spurious negative association due to smaller studies using more informative imaging modalities. I excluded the review study by Cuingnet et al. [88], as only studies advancing new learners on the basis of performance will be liable to selection bias. I measured sample size using only the relevant subjects (AD and HC subjects in AD detection, and MCI subjects in MCI prognosis). Where a range of performance results are reported, I have selected the mean as the reported performance of the study to avoid any bias not implied by its authors. In total, this lead to 55 performance estimates for AD classification and 15 for MCI prognosis.

The collected performance results are presented in figure 5.11. A negative association between sample sizes and reported performance is readily apparent in both classification tasks. For both of these, I computed the Spearman's rank correlation coefficient. I also computed an associated p value to assess the evidence against a null hypothesis of zero expected rank correlation, as would be done in a one-sided test. The use of rank correlation rather than product

[7] https://competitions.codalab.org/competitions/1471

moment correlation to generate p values means that no assumption of a joint Gaussian distribution is required. The p values for the AD detection and MCI prognosis tasks were 0.0089 and 0.0707 respectively. I note that the evidence against the expected positive association is stronger than the evidence against a zero association as demonstrated by the p values. I present this as evidence suggestive of a negative association between sample size and reported performance in the literature, consistent with the presence of significant selection bias.

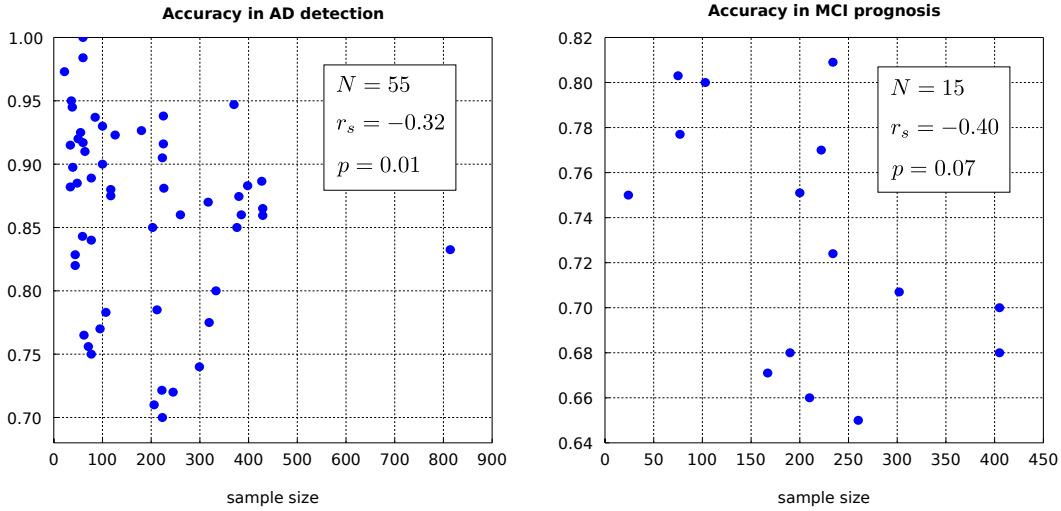

Figure 5.11: Summary of sample sizes and balanced accuracies from papers considered in [29] using structural MRI alone. The selection of the studies and the production of performance measures is described in section 5.4.2. The variable r_s denotes Spearman's rank correlation, and the p values are derived from a one-sided test against a non-negative rank association. N denotes the number of studies included in a tasks' plot.

5.4.3 Reducing selection bias

Selection bias is an inevitable consequence of the use of CV results to refine learners. Though they cannot eliminate it, researchers may take precautions to reduce its degree and mitigate its effects. Broadly speaking there are two ways to do this: reducing the variance of performance estimation and reducing the number of learner specifications considered. Variance reduction should also improve the ability of researchers to identify superior learner, as demonstrated by the effect of sample size in figure 5.10.

5.4.3.1 Variance reduction

The most simple way to reduce the variance of performance estimation is to maximise the full sample size. While, in some cases, this may be done by including a more diverse set of images [43], there may be little scope to do so otherwise. Another, possibly more achievable, method may be the choice of a low variance CV strategy. Where using ExKcv, higher values of E can be used to this effect (see figure 5.6). Where computational power is limited, using a lower value of K can provide a variance (and bias) reduction without an increase in

computational effort (see figure 5.6). While high K strategies such as leave-one-out cross validation (LOOCV) can provide higher expected performance, they may actually hinder accurate performance estimation and correct learner ranking through their limited capacity for variance reduction with increasing E (see [35] and section 6.5). As noted by Cawley and Talbot in [143], the variance of a CV strategy is at least as important for learner selection as the 'bias' associated with a reduction in training set sizes. Though LOOCV may be recommended for use with smaller sample sizes, that combination will produce the greatest possible potential for selection bias.

5.4.3.2 Specification number reduction

There are several ways that researchers may reduce the number of specifications considered. Additional processing steps increase the number of possible learner specifications exponentially (as can clearly be seen in the construction of the experiments of this chapter), and provides more scope for selection bias. Even in a pool of learners that are ill-suited to a prediction task, there are likely to be some impressive results in a finite sample if that pool is large enough. Researchers can avoid selection bias by restricting their consideration to a smaller pool of learners justified by domain specific knowledge, rather than using a more naive search through the combinatorial space of specifications.

Crucially, the amount of 'optimisation' that can productively be conducted will depend on the amount of available data. Large samples that comprise hundreds of subjects (such as the one collected for this study) may allow researchers to usefully refine and develop learner specifications. Conversely, efforts to fine-tune learners using a sample of only 50 subjects are likely to yield little real out-of-sample performance improvement (see figure 5.8). 'Premature' learner optimisation conducted before large samples are available is likely to be misleading and ultimately of limited utility.

5.4.3.3 A bias free validation strategy

In the context of parameter optimisation, the problem of selection bias has already been recognised and dealt with by the ANA community. The solution they selected was to encapsulate the selection of parameters into learner specification through nested CV. Cross validation performance estimates then apply to 'a learner with its parameter selected using CV in the training set' rather than to 'a learner with its parameter set to X'. This subtle change in interpretation is not thought to be a problem.

The same solution could be extended to the problem of specification selection [143, section 5.1]; that is, learner specification could be treated as a parameter to be selected through nested CV. The resulting 'meta-learner' is one that uses the training set to select from the set

of all specifications a research group wishes to consider. Now, the resulting performance estimate should be treated as that of 'the learner that appears to be the best based on performance estimated through CV on the training set' rather than 'learner specification X'. For a research group reporting their results this way, this produces an unbiased estimate of 'what is the best that we can do' at the cost of clouding exactly 'how can we achieve best results'.

5.4.4 Role of transparent reporting

By reporting all attempts at learner improvement, rather than just the successful ones, challenges such as CADDementia [25] make it easier to identify scenarios where selection bias is likely to be significant. The greater the role that random effects play in determining a performance ranking, the greater the selection bias associated with the best performing learner will be. When the differences between the true performances of the learners are small compared the standard deviation of the performance estimation, the distribution of observed performance results will appear in a single roughly Gaussian cluster. In these circumstances, it is reasonable to expect that selection bias will be high. Conversely, when a single unimodal distribution describes the observed performance measurement poorly, it more is likely that measurement noise plays a less important role in determining the empirical performance ranking, so selection bias should be lower relative to the observed variability.

5.4.5 Limitations of this study

The quantitative description of selection bias here is a function of the exact learner specifications considered, and the experiments here are not a precise simulation of the selection process enacted by the research community. This would be impossible, as the number of variety of unpublished experiments cannot be known. The measures of bias presented here are indicative rather than exact. In particular, k_{select}, the X-axis of figure 5.8, should not be taken as a description of the exact *number* of learners considered for selection in a general context, but as a description of their *diversity*[8]. I imagine that real research practices involve a more diverse set of learners than those presented here, as many of mine are very closely related. Due to constraints on running and implementation time, some learner specifications considered here are relatively simple. While these include many with performance much lower than those commonly reported in the literature, I believe their presence does reflect the situation in real research (where low performance measurements may go unreported). The inclusion of the learners with widely differing performance actually acts to reduce the role which random effects play in the

[8] I note that when, in error, I originally included a set of 24 extra feature sets that were near duplicates of the kernel feature sets presented, this had the effect of rescaling the curves of figure 5.8 while almost perfectly preserving their shape.

empirical rankings, so I do not think that it has lead me to overestimate the importance of bias.

Finally, I did not provide an uncertainty estimation with my estimates of selection bias. It is assumed that the noise in the estimation of these measures is of an order no greater than the deviation from the values taken in real research caused by differences in the selection process and the learner specifications considered. One could nest the bias measuring subsampling experiment within a similar subsampling experiment to measure the variance of all relevant quantities empirically [35]. However, this would limit the sample sizes for which bias can be estimated to half those used here. It would also greatly increase the computation complexity of the experiment.

5.5 Conclusion

My results demonstrate that selection bias is a potentially significant concern for the AD classification research field, as it can account for an appreciable fraction of the apparent improvement in performance associated with the optimisation of learner specifications. The implied level of selection bias can explain the lower than expected correlation between sample size and reported performance in the AD classification literature, and is consistent with bias observations in classification challenges.

The crucial determinants of selection bias are the number/diversity of learners considered for selection and the variance in performance estimation. Variance can be controlled by maximising sample size and using low variance CV strategies such as RKCV with low values of K and high numbers of KCV repetitions. I urge against the premature optimisation of learners before sufficiently large samples are available, and I call for caution in the interpretation of results from small-sample studies.

While this study focused on AD classification in the ADNI dataset, selection bias in any ANA problem that receives extensive research attention is likely to behave in a way that is qualitatively similar.

Part III

Cross validation strategies

Chapter 6

Better cross validations strategies

This chapter is dedicated to the identification of superior cross validation (CV) strategies for use in AD ANA. I shall begin the chapter with several short analyses on aspects of CV strategies which should provide insight into why some strategies are to be preferred over others. I shall then consider various novel strategies that may be used to improve on current practice, which is mostly K-fold cross validation (KCV) and repeated K-fold cross validation (RKCV). The first of these is extended K-fold cross validation (EKCV), a new CV strategy that I have developed to extend KCV to use a wider range of training set sizes. EKCV is particularly useful in experiments where the training set size must be precisely controlled. After EKCV, I shall then discuss the optimal selection of K in KCV and RKCV, as well as various uncommon alternative CV strategies from the literature, including the .632+ bootstrap. The chapter concludes with a series of strategy recommendations for AD ANA researchers.

6.1 What makes a cross validation strategy desirable?

The key factors that make a CV strategy desirable are bias, variance and computational cost.

The expectation and (training set) bias of a strategy are essentially defined by the size and subpopulation composition of the training sets, and are identical among all **commensurate** strategies (see section 3.5.1). Recall that a CV strategy is specified by a random design $\mathbf{I} = \langle I_r \rangle_{r=1}^{R}$. In a non-stratified context, all strategies where $|I_r| = m$ for all r are commensurate.

For a given set of learners, computational cost is largely determined by R, the number of train-test experiments.

Among commensurate strategies, variance can be minimised by using additional train-test experiments. In RKCV or repeated hold-out cross validation (RHOCV), this can be achieved by simply increasing the parameter E, describing the number of random KCV or single hold-out cross validation (SHOCV) experiments used. As discussed in section 3.5.2, the variance can be divided into two parts in this case: one irreducible part that is unaffected by E and one reducible

part proportional to E^{-1}. The lowest possible variance among all strategies using m items in the training set is achieved by leave-p-out cross validation (LPOCV), which uses all possible compliant train-test splits.

Where computational resources are limited, variance can be reduced by choosing a strategy that is more **efficient**. This is achieved by selecting component train-test experiments such that the correlation between their results in reduced. Efficiency is a highly desirable attribute for a CV strategy, and there will rarely be a reason to choose the least efficient of two commensurate strategies.

In some cases, there may be a trade-off between bias and variance. In that case, the ideal choice of strategy will depend on the task being considered.

6.1.1 Learner performance estimation

In learner performance estimation, bias b and variance v contribute to the squared error of a CV performance estimator in the standard formula $b^2 + v$. In AD ANA, low error performance estimation is desirable because it allows researchers to assess the potential benefit associated with new diagnostic systems. In this context, because it is most important that a certain minimum standard is reached, one may potentially be less concerned by negative bias (associated with limited training set sizes) than one would be with a positive bias.

6.1.2 Learner selection

Instead of directly estimating the performance of a single learner in an anticipated future application, one may be primarily interested in identifying which of two or more candidate learners will have the highest performance there. In this case, it is most important to estimate the *ranking* of the learner performances, rather than their specific values. This is one of the key goals of AD ANA, where it corresponds to the identification of the best machine learning pipelines for diagnostic decision making and clinical trial enrichment.

As Cawley and Talbot note [143], while low bias is often cited as an important concern in learner selection, low variance is at least as important. As long as learner rankings are stable over the relevant range of training set sizes, it may not matter that the size used in CV is different to the one anticipated in a future application.

The stability of learner performance rankings over a large range of training set sizes is particularly plausible in ANA, where learner performance rankings are often explained in terms of the fundamental relationships between the learners and problem under study that are unlikely to change rapidly with training set size. For instance, it may be said that ‘imposing sparsity is beneficial because only a small subset of the features is relevant’ or that ‘automatic weight

selection is better able to exploit multi-modal data'. Not only is ranking stability implied, it is also assumed; because the precise training set size used in the future application are unknown, any ranking of learners before that application exists requires learner performance rankings to be stable over an anticipated range.

As discussed in chapter 5, in a non-ideal setting where not all results are reported, variance is also an important determinant of selection bias.

Summary

Low training set bias and low variance are desirable aspects of a CV strategy. Both are important in performance estimation, but variance is likely to be more important in learner selection.

High efficiency is particularly desirable, as it allows low variance to be achieved without additional computation.

6.2 Factors determining variance

Recall that the performance measurement of a CV strategy with a design $\mathbf{I}$ is given

$$\Gamma(u,\mathbf{D},\mathbf{I}) = \frac{1}{R}\sum_{r=1}^{R}\gamma(u,\mathbf{G}_r,\mathbf{H}_r)\,. \tag{6.1}$$

Where all the $\gamma(u,\mathbf{G}_r,\mathbf{H}_r)$ have the same marginal distribution, the variance of $\Gamma(u,\mathbf{D},\mathbf{I})$ is

$$\mathbb{V}\mathrm{ar}\big[\,\Gamma(u,\mathbf{D},\mathbf{I})\,\big] = R^{-2}\sum_{1\le r\le R}\mathbb{V}\mathrm{ar}\big[\,\gamma(u,\mathbf{G}_r,\mathbf{H}_r)\,\big] + 2R^{-2}\sum_{1\le r'<r\le R}\mathbb{C}\mathrm{ov}\big[\,\gamma(u,\mathbf{G}_{r'},\mathbf{H}_{r'}),\gamma(u,\mathbf{G}_r,\mathbf{H}_r)\,\big]. \tag{6.2}$$

The first term on the right hand side is identical for all strategies using a given R train-test experiments and training set size m. The second term is proportional to the average correlation between test experiments. It is this that determines the relative efficiencies of commensurate strategies with identical R.

6.2.1 With sequential random experiments

In RHOCV and RKCV, the final performance estimate is the mean of E exchangeable performance estimates corresponding to individual SHOCV or KCV repetitions using randomly generated partitions. Let σ^2 denote the variance of the individual estimates, and ρ denote the covariance between all pairs of them. The variance of the final performance estimate is given

$$\sigma^2\Big(\frac{1}{E}+\frac{E-1}{E}\rho\Big) = \sigma^2\rho + \sigma^2\frac{1-\rho}{E} \tag{6.3}$$

In this case, the left and right terms correspond the irreducible and reducible variance components respectively. These may be compared to the corresponding expressions in equation (3.21)

of section 3.5.2.

6.2.2 Equal use strategies: implications of the fixed predictor model

Under the fixed predictor model, all predictors constructed by a learner in a CV experiment are equal to some constant t. In this case, the performance on the ith item of $\mathbf{D} = \langle D_i \rangle_{i=1}^{l}$ is the random variable $Q_i = (t(X_i), Y_i)$, where $D_i = (X_i, Y_i)$. The Q_i are assumed to be i.i.d. with some shared mean and variance σ^2.

Consider a CV experiment with a design $\mathbf{I} = \langle I_r \rangle_{r=1}^{R}$ specifying the training sets of R component train-test experiments. The testing sets are specified by the index subset $J_r = \{1,2,...,l\} \setminus I_r$. For all r, $|J_r| = n$ and $|I_r| = m$. In this case, the final performance estimate is given

$$\frac{1}{Rn} \sum_{r=1}^{R} \sum_{\iota \in J_r} Q_\iota = \frac{1}{Rn} \sum_{\iota=1}^{l} a_\iota Q_\iota \,, \tag{6.4}$$

where a_i is the **testing inclusion count** of item i, defined as the number of times it has been included in a testing set. This may be expressed as

$$a_i = \sum_{i=1}^{l} \mathbf{1}_{i \in J_r} \,,$$

in which $\mathbf{1}_{i \in J_r}$ represents the indicator function taking the value 1 when $i \in J_r$ and 0 otherwise. By definition,

$$\sum_{i=1}^{l} a_i = Rn \,. \tag{6.5}$$

By the standard formula for the variance of a weighted sum of random variables, the variance of the CV performance estimate defined in equation (6.4) is

$$\frac{\sigma^2}{(Rn)^2} \sum_{i=1}^{l} a_i^2 \,. \tag{6.6}$$

Because the design $\mathbf{I}$ is randomly generated, the a_i may be regarded as random variables. To compare strategies with different random designs, one must consider the expected variance,

$$\frac{\sigma^2}{(Rn)^2} \sum_{i=1}^{l} \mathbb{E}[a_i^2] \,. \tag{6.7}$$

To produce a design with minimal variance, one must minimise the quantity of equation (6.6) subject to the constraint of equation (6.5). When the a_i share a single marginal distribution, this is equivalent to minimising $\mathbb{E}[a_i]$. This will occur when the testing inclusion

counts are as equal as possible[1] This condition is met in KCV, RKCV and LPOCV, which may be called **equal use strategies**. In all of these, $Rn \bmod l = 0$, and $a_i = Rn/l$ for all i. *Under the fixed predictor model, all equal use strategies have the minimum possible variance*, which by expression (6.6) must be equal to σ^2/l.

Equal use strategies may be contrasted with RHOCV, where $a_i \sim Bin(R, n/l)$. Because, for any binomial random variable $X \sim Bin(n,p)$,

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[X]^2 + \mathbb{V}\text{ar}[X] \\ &= (np)^2 + np(1-p)\,,\end{aligned}$$

in RHOCV,

$$\mathbb{E}[a_i^2] = \left(\frac{Rn}{l}\right)^2 + R\frac{n(l-n)}{l^2}\,. \tag{6.8}$$

Combining this result with the formula for the variance in equation (6.6) gives the expected variance of RHOCV as

$$\frac{\sigma^2}{l}\left(1 + \frac{l-n}{Rn}\right), \tag{6.9}$$

which is strictly greater than the variance seen in an equal use strategy with the same R and l. *Thus, KCV, RKCV and LPOCV strategies are all more efficient than commensurate RHOCV using the same number of train-test cycles*.

6.2.2.1 Accuracy of fixed predictor model variance predictions

While the fixed predictor model is able to explain the efficiency benefit of equal item use, it also predicts that all equal use strategies will produce estimators with the same variance. It is therefore unable to explain why RKCV, or indeed LPOCV, should produce lower variance estimators than KCV. In general, one should expect the fixed predictor model to be most accurate in cases where predictor selection is 'stable', and there is little random variability associated with a training set. This will tend to occur in cases where the number of training items is high relative to the number of parameters. Conversely, it will be least useful in cases where predictor selection is 'unstable'.

Because they are both caused by deviations from the fixed predictor model, variance reduction with increasing E in RKCV and type I error rate inflation associated the problem of dependency are likely to occur in the same experimental contexts.

[1] $Rn \bmod l$ of them will have a value of $\lfloor Rn/l \rfloor + 1$, and the remaining $l - Rn \bmod l$ will have a value of $\lfloor Rn/l \rfloor$.

Summary

Equal testing use is an important determinant of the efficiency of a CV strategy. This expected even under the fixed predictor model.

6.3 Subpopulation stratification

In this section, I will review the effects of subpopulation stratification in CV. Specifically, I shall consider its effects through actions on the training and testing sets of a CV experiment respectively.

6.3.1 Effect in testing sets

When a random testing set is used to measure the performance of a fixed predictor in a non-stratified context, the relative fractions of the items in the testing set from each of the subpopulations will vary. This will introduce some additional random variation into the final performance measure. If one uses stratification to keep the relative contribution of the different subpopulations to the testing set the same as their relative contributions to the full population, then this will reduce the variance of the final performance estimate without affecting its expectation. A proof for this is included in appendix B.

The same variance reduction due a more balanced testing set should be expected when using stratification in SHOCV. However, in KCV and RKCV, all items will contribute equally to the final performance estimate even in the absence of stratification. Under the fixed predictor model, the relative contribution of the items to the final estimate should be all that is needed to determine the bias and variance of a CV strategy, so stratification would have no effect. In KCV-like strategies, the more important effect of stratification is related to the training sets.

6.3.2 Effect in training sets

As described in section 4.2.2, the size of a randomly generated training set is an important determinant of the distribution of the predictors a learner will select on it. Larger training sets provide better 'coverage' of the distribution of the items, and so convey more information about it. This will generally lead to a more effective selection of predictor. When stratification is used to produce a training set with a representative balance of items from different subpopulations, it will tend to have a more even coverage of the full population distribution, and so predictors selected on it will tend to have higher performances. This is particularly important in classification settings, where the balance of the subpopulations determines the prior probability a predictor should give to each label in the absence of discriminative information in the features. Where one expects the unseen dataset on which one's ANA method will be trained in a future application to be stratified to increase performance, one should also use stratification in

preliminary research, as this will reduce bias [159].

Not only will the performance of a predictor selected on a stratified training set tend to be higher than one selected on a non-stratified training set of the same size, but it will also tend to be more stable, as the degree of coverage offered by a stratified training set will be more consistent. It is likely this effect that is important in giving stratified KCV-like procedures lower variances than their non-stratified equivalents [159].

6.3.3 Reduced number of splits

While using subpopulation stratification will tend to reduce the variance of a CV performance estimate, when using strategies with very high numbers of train-test splits (e.g., LPOCV), it could conceivably raise it. This is because the use of stratification reduces the number of possible train-test partitions of a sample, meaning that their are fewer example predictors to incorporate into the final performance estimate. This is unlikely to occur in AD ANA.

6.3.4 Stratification-like procedures for K-fold cross validation

For KCV and RKCV in particular, there are two stratification-like methods that make the distribution of the training set more representative. These are distribution balanced stratified cross validation (DB-SCV), presented in [160], and distribution optimally balanced stratified cross validation (DOB-SCV), presented in [161]. Both of them use a measure of distance between the feature space representations of items to try to place similar items in different disjoint subsets of a KCV partition. This achieves a similar balancing effect to subpopulation stratification without the need to divide the sample into disjoint subsets. Indeed, both methods can be used simultaneously with subpopulation stratification. Of the two, DOB-SCV appears to have better performance in empirical studies [161].

Both DB-SCV and DOB-SCV are implemented similarly. Briefly, to perform DOB-SCV, one iterates through each subpopulation and selects one of its items in $\mathbf{D}$ at random. One then selects the $K-1$ items of the same subpopulation that are nearest to it. Each of the resulting K items are placed in one of the K disjoint used to define a KCV experiment. The selected items are removed from further consideration, and the process is repeated until all items have been distributed to the K subsets. For both methods, the selection process means that the items of the training sets are no longer independently generated from well defined populations of items. Both the sizes of l and of m will affect the distribution of the training sets, with larger samples allowing a more representative selection of items for a given m. This may complicate interpretation.

A related method to ensure representative training and testing sets in KCV partitions is the

one of Zhang and Qian [162], which uses latin hypercube sampling to that all training sets have similar marginal distributions across each feature. This method is not applicable in research on real data, as it requires the items to be computer generated from a known distribution.

Summary

Stratification is an important tool for the reduction of bias and variance in a CV strategy, particularly in classification tasks. There are other stratification-like procedures that may be used in KCV and RKCV specifically.

6.4 Extended K-fold cross validation

This section describes a new CV strategy I have developed to retain the equal use characteristic of KCV and RKCV strategies while relaxing the restrictions on the size of the training sets that may be used.

6.4.1 Motivation

In many contexts, one may wish to study the effect of the training set size and/or class composition on the performance of a learner. (For instance, one may wish to see what training sample size would be sufficient in a future application.) In theses cases, it is necessary to construct a series of cross validation experiments in which m can be precisely controlled. At the same time, one will wish to use an efficient low variance strategy with a practically achievable number of train-test repetitions (precluding not LPOCV). While KCV and RKCV are both relatively efficient due to their equal use of items for testing (see section 6.2.2), they restrict an experimenter to values of m such that, for some integer K, $m \approx l - l/K$. If K does not divide l, then a mixture of training set sizes $l - \lfloor l/K \rfloor$ and $l - \lceil l/K \rceil$ are then used. One cannot use training set sizes below $l/2$ and one cannot freely vary m with great resolution. For instance, no values of m between $l/2$ and $2l/3$ are possible.

Extended K-fold cross validation (EKCV) extends KCV to allow for a greater range of training set sizes while retaining the equal testing use constraint. It has been applied to study the effect of training set size on learner performance in [163] and appears in the later work of this thesis.

6.4.2 Implementation

EKCV uses a greedy selection of index subsets to minimise the variance measure of equation (6.6). To do this, one keeps a record of the testing inclusion counts of each of the items of a full dataset $\mathbf{D}$. These are all initially set to zero. After r train-test experiments, all the a_i will have values $\lfloor rn/l \rfloor$ and $\lceil rn/l \rceil$ (though these may be the same). When using stratification, one

simply draws the indices for each subpopulation separately, and then combines the subpopulation index sets. This is achieved with the algorithm describe in figure 6.1. If $m < n$, it may be

```
set all item inclusion counts to zero
for each train-test experiment index r ∈ {1,...,R} do
    for each stratification group do
        add all group items to candidate list
        remaining number of group items needed in testing set is n_remain = n
        while n_remain > 0 do
            select an item at random from set in the candidate list with the lowest
             inclusion count
            add selected item index to tests set J_r
            and remove it from candidate list
            increment that items inclusion count
            decrement n_remain
        end
    end
    select I_r as {1,2,...,l} \ J_r
end
```

Figure 6.1: Description of the subset selection algorithm used for EKCV

more efficient to reverse the roles of I_r and J_r, and to instead minimise the training inclusion counts defined by the appearance of items in the I_r rather than the J_r. This has the effect of jointly minimising the testing inclusion counts. Generally, R and m should be chosen such that $Rn \bmod l = 0$, so that all items are used equally for testing. Note that this is always the case when $R \bmod l = 0$.

6.4.3 Characteristics and reduction to RKCV

When m and R are chosen such that $Kn \bmod l$ and $R \bmod K = 0$, EKCV will reduce to RKCV with K fold and R/K repetitions. This means that, by definition, *EKCV has identical training set bias and variance characteristics to RKCV when these parameter choices are made*. The effect of varying the parameters of EKCV/RKCV is explored in the next section.

6.5 What is the correct choice of K/m in KCV?

The correct choice of K is KCV or RKCV is of great practical importance, as it will determine both bias and variance. This issue is considered in Kohavi's study of 1995 [159]. In it, after a set of empirical experiments, he decides that intermediate values (between 2 and l) are better, as both low and high values of K lead to high variance. The high variance of leave-one-out cross validation LOOCV in particular has been noted by both me [35] and others elsewhere [122, section 7.10].

The choice of K in KCV/RKCV actually has two important effects. First, it changes the

training set size m used in the component train-test experiments. Second, it changes the number of those experiments, and thus also the computation cost of the CV. I note, as Kim does [158], that it is most informative to compare CVs strategies with equal computation costs. This type comparison will be most useful to future researcher who must decide on a CV strategy where there are limited computation resources. Therefore, when selecting a KCV or RKCV-like strategy, one should not ask what the right choice of K is. Rather, one should ask what the right choice of m is for a fixed number of train-test repetitions. If one is interested estimating the learner performance associated with l items in the training set, there is a bias-variance trade-off in the choice of m in EKCV/RKCV. In general, lower values of m appear to provide lower variance estimators, but are associated with lower learner performances.

Figure 6.2 presents the results of a small experiment I conducted to illustrate this point. I created a synthetic binary classification problem based on two 9-dimensional Gaussian distributions with a covariance structure specified by the identity matrix and a separation of 1.3. I used the following classification algorithms:

1. the nearest centroid (NC) algorithm, which should be well suited for this problem,
2. KNN with $K = 1$,
3. KNN with $K = 3$,
4. KNN with $K = 5$, and
5. the C4.5 decision tree algorithm.

Varying m between 4 and 76 in steps of 4, I then performed class stratification EKCV with $R = 40$ on 10000 independently generated samples of 80 (40 of each type) items to measure the mean and variance of the resulting accuracy estimates. I also measured the learner performances associated with a training set of 80 items. With the mean and variance estimates, I was then also able to estimate the mean squared error associated with EKCV using each choice of M.

6.5.1 Discussion

As figure 6.2 demonstrates, even when one's goal is to estimate the learner performance associated with a training set that is the same size as the full available sample (i.e., l), it may still be better to use a smaller training set size. This is because of the increased training set bias is more than compensated for by the reduced variance. Of course, different learning problems will have different bias and variance responses to changes in m, so there will be no universally optimal choice. As demonstrated by the differences between the algorithms, the steeper the change in

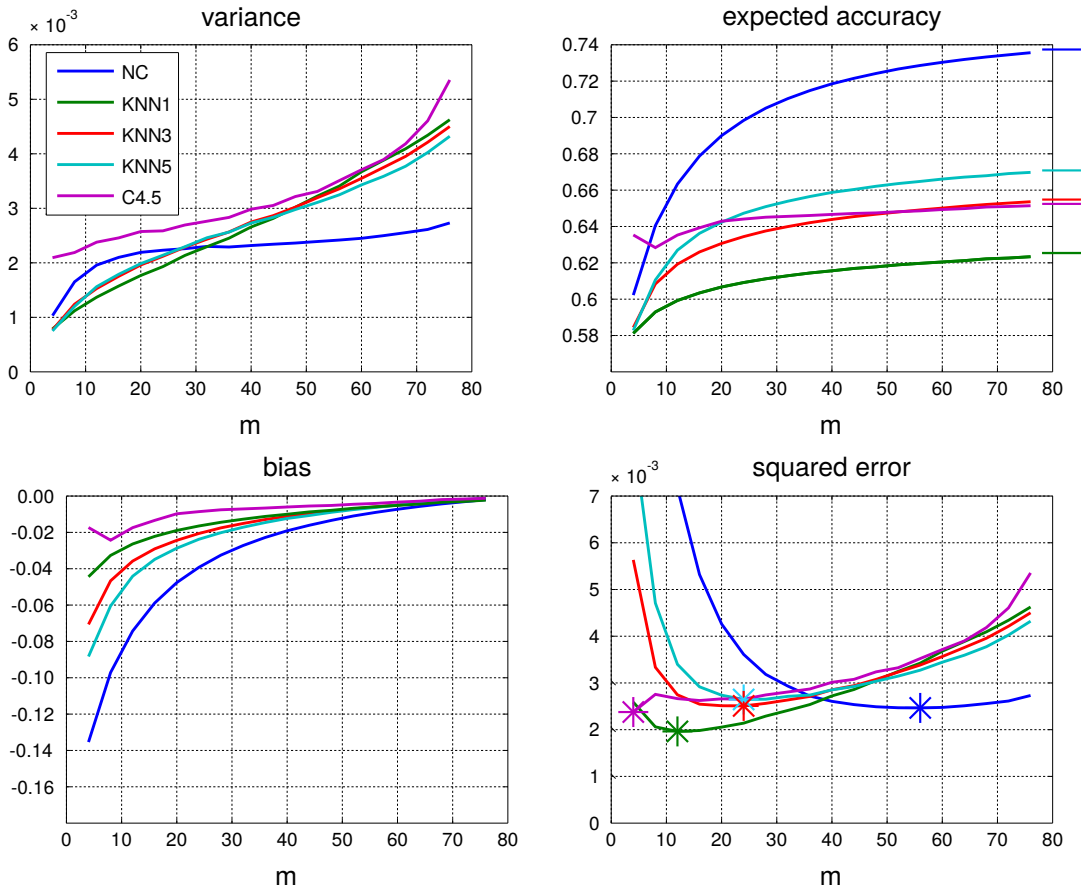

Figure 6.2: Illustration of the bias-variance trade-off associated with training set size selection in EKCV. Variance, expectation, bias, and squared error curves are presented for the estimation of the full sample learner performance in the synthetic classification problems described in section 6.5. The full sample learner performances are represented as bars on the right of the expectation plot. The stars represent the minimum achievable squared error.

the learner performance with m, the more it is worth using a higher value of m. However, even for the steepest curve, the optimal choice of curve was $m \approx 2l/3$.

As discussed in section 6.1.2, when one wishes to rank two learner performances or estimate the difference between them, the issue of bias should be less of a concern than it is in single learner performance estimation. Consequentially, there is a particular motivation for smaller training set sizes in learner selection.

6.6 Uncommon cross validation strategies

In this section, I shall describe some other CV strategies that have been proposed as more efficient alternatives to some of those already described and discuss whether these hold potential to improve validation in ANA.

6.6.1 Repeated learning testing

In their review of CV strategies, Arlot and Celisse describe a strategy called repeated learning testing (RLT) in which train-test experiments are sequentially selected at random as in RHOCV, but where a record of these is kept to ensure that no one experiment is repeated [36, section 4]. Where m items out of l are used for training, RLT will converge to LPOCV as the number of train-test experiments reaches $\binom{l}{m}$. In fact, the paper the authors cite as the source of RLT is actually describing RHOCV [164], making RLT an accidental invention of Arlot and Celisse. Nevertheless, I think it is worth giving the idea some consideration.

When a small number of train-test experiments are conducted, the chance of 'collisions' (repetitions of a train-test experiment) occurring is low, and RLT will be essentially identical to RHOCV. The number of train-test experiments required to make one or more collisions occur with a probability of at least $1/2$ is generally very high. In fact, it is asymptotically proportional to $\binom{l}{m}^{1/2}$. To give an example, if $l = 50$ and $m = 40$, 119334 train-test experiments are required for this to occur. If such a large number of train-test experiments is conducted, then it becomes necessary to keep some record of which of the very many partitions (train-test experiments) have already occurred, and to check against this repeatedly. This is a potentially very demanding task in terms of memory and computation, and is likely to offer only limited rewards; as RLT does not enforce equal use of items in training or testing (see section 6.2.2), it is likely to have higher variance than, say, RKCV using the same number of train-test experiments.

6.6.2 The .632+ bootstrap

The .632+ bootstrap is a CV strategy designed for classification tasks by Efron and Tibshirani in [165]. Unlike the other strategies discussed, the training sets used in .632+ bootstrap may include multiple copies of a single items index in $\mathbf{D}$. Each of the training sets $\mathbf{G}_r$ used is produced by sampling with repetition from $\mathbf{D}$; in this case, I_r must be treated as a sequence of indices rather than a set, and its elements are random integers uniformly distributed between 1 and l. The testing set $\mathbf{H}_r$ is specified by a testing index set J_r containing one instance of all those indices in the set $\{1,2,...,l\}$ that do not appear in I_r. The final performance estimate provided by .632+ validation is not the standard average over testing sets, as a corrective function is applied to compensate for the duplication of items. This function is intended to produce an unbiased estimate of the learner performance associated with a training set of l items.

The .632+ bootstrap has been shown to have desirable low variance in empirical studies [158], being in some cases more precise than RHOCV and RKCV using the same number of train-test experiments. Despite this, it suffers from a high level of training set bias [158]. The corrective function intended to eliminate bias is based on a loose argument about inter-

point distances. It has been criticised as having "nearly no theoretical justification" [36] and is even described by its authors as "really quite rough" [165]. While it may not be a great problem that the .632+ is biased, it is difficult to anticipate what that bias will be. This will make interpretation difficult. The expected values of a estimate produced by the .632+ bootstrap does not correspond to any learner performance as defined by a train-test experiment with i.i.d. items. I find this reason sufficient to recommend against the use of the .632+ bootstrap in ANA.

6.6.3 Balanced incomplete cross validation

As described in sections 6.2.2, where design selection is constrained to ensure that all items are use equally for testing, this should help produce a lower variance performance estimates. It is possible to further constrain the selection of train-test partitions to ensure that all *pairs of items* are used equally too, and this may have some benefit in terms of efficiency by reducing the average covariance between component train-test experiment results. The original author to propose this was Shao in 1993 [166]. He called the resulting strategy balanced incomplete cross validation (BICV). Recently, Fuchs, Krautenbacher and others have revisited this idea as a way to produce a lower variance strategy than RKCV using the same level of computational effort [135, 167]. Unfortunately, BICV may not be used with stratification, and it places strict constraints on the training set size m and train-test experiment number R. This makes it currently impractical for AD ANA. However, future extensions may relax these restrictions, allowing a BICV-like strategy to offer a more efficient alternative to RKCV. This idea will be visited in more detail in the next chapter.

6.6.4 Short-cut CV

There are some specific learning algorithms for which 'short-cuts' exist for LPOCV or LOOCV. These short-cuts allow one to implicitly perform exhaustive CV with a greatly reduced computational cost. Examples include K nearest neighbour methods [119], kernel discriminant classification [168], and some formulations of SVM [169]. When one is interested in a single learner for which a short-cut is available, one may choose exhaustive CV over other strategies that would otherwise be preferred. Of course, these methods can only be applied if there are no pre-processing steps that use the labels. Unfortunately, short-cuts normally place heavy restrictions on the type of CV used. Because ANA requires a comparison of a great many varied learners, short-cut validation will be of limited use.

6.7 Strategy recommendations for AD ANA

Based on my review of the literature and my discussions with researchers, I find that too much importance is attached to training set bias in AD ANA, while too little is attached to variance.

I note that this pattern that may be present in supervised learning literature in general [143]. In AD ANA specifically, there is another reason why bias may be less relatively important; as discussed in section 6.1.2, learner performance rankings are expected to be (and described as being) stable over a reasonable range of training set sizes. This should make the small bias associated with, say, a mild reduction in training set sizes, relatively unimportant for the comparison of methods in a single study.

As discussed in section 3.8.3, many studies use only one KCV or SHOCV repetition where they could use a larger number. Where it is possible, they should do this, as that will reduce the variance of the resulting performance estimation, leading to better performance estimation and learner selection. While ANA methods may be computationally intensive, much of their computational load is associated with the unsupervised image processing steps (e.g., registration, segmentation) that do not need to be repeated in sequential CV experiments. In some cases, the choice of KCV or SHOCV over lower variance RKCV may be justified by a desire to avoid more severe dependencies in any statistical analysis. As will be discussed in chapter 10, this motivates the development of analysis techniques that can be used with low variance high R CV strategies.

Another feature of CV in AD ANA discussed in section 3.8.3 is the use of strategies where a relatively high fraction of the sample is used for training in each train-test experiment (e.g., LOOCV, KCV with $K = 10$). The choice of high variance, high m strategies may be due to concerns about training set bias, but it may also be due to perverse incentives created by a lack of training set size consideration in performance comparison over studies (see section 3.8.2.1). When learners must compete based on their reported performance without reference to m, researchers wishing to demonstrate the utility of their own method will wish to maximise the training sets, despite the associated loss in precision. Lower training sizes, such as $m \approx 2l/3$, should be used in CV strategies with high numbers of train-test repetitions. As discussed in section 6.5, this will often produce a better procedure even when the sole goal is the estimation of a learner performance at a large training set sizes, and it will certainly do so when the goal is learner selection. I reiterate the advice of [143]: variance is at least as important as bias in this task.

AD ANA researchers should continue to use equal testing use strategies, as these will be somewhat more efficient. This should be RKCV or EKCV with some high number of train-test experiments whenever computational feasible. Class stratification should continue to be used, as the more effective and stable selection of predictors it offers can reduce bias and variance in performance estimation with little negative effect. DOB-SCV might also be used to increase

precision and reduce bias. If it is, researchers should remember that larger samples will produce better balanced training sets for a given m, so results in different studies may not be comparable even if the training set sizes used are the same. I did not find that any of the uncommon strategies discussed in section 6.6 offer much potential for better practice in AD ANA in their current form, though a modified form of BICV may offer a more efficient alternative to RKCV and EKCV. This idea will be explored in the following chapter.

Chapter 7

Balanced incomplete cross validation

This chapter is devoted to balanced incomplete cross validation (BICV), a form of CV that may be more efficient than KCV and related variants. I shall begin by introducing block designs in section 7.1, and then discuss how these can be usefully applied to CV. I shall describe how designs can be used to increase efficiency, with particular focus on the recent work of Fuchs and Krautenbacher in [135]. In section 7.2, I shall introduce approximately balanced cross validation (ABCV), a new strategy I have developed as an effort to extend BICV to a wider range of experimental settings while retaining some of its superior efficiency. In section 7.3, I shall then describe some preliminary validation experiments on real and simulated data, before presenting concluding remarks in section 7.4.

7.1 Block designs for cross validation

7.1.1 Introducing block designs

A block design is a sequence $\mathbf{I} = \langle I_r \rangle_{r=1}^{R}$ of R subsets of a set of the integers $\{1,2,...,l\}$. These subsets are called blocks. The order of the elements of $\mathbf{I}$ is not normally considered important, so designs that are permutations of each other are normally considered equivalent. As shown in chapters 3 and 6, block designs can be used to describe CV. Of particular interest in CV are uniform designs where all blocks have a fixed size m. Every design $\mathbf{I}$ has a **complement** design $\mathbf{J} = \langle J_r \rangle_{r=1}^{R}$ specified $J_r := \{1,2,...,l\} \setminus I_r$. In CV, where a design $\mathbf{I}$ specifies the training indices, the complement design $\mathbf{J}$ specifies the testing indices.

The number of times that a subset $\theta \subset \{1,2,...,l\}$ appears in the blocks may be called the **inclusion count** of that subset. Formally, the inclusion count of θ may be defined

$$a_\theta = \sum_{r=1}^{l} \mathbf{1}_{\theta \subset I_r}, \tag{7.1}$$

Using this, one can define the inclusion count of a single item with a set of the form $\{i\}$, a pair

of items with the set of the form $\{i, j\}$, and so on. A b-design is a design in which the inclusion counts of all $\theta \subset \{1,2,...,l\}|$ such that $|\theta| = b$ take the shared value λ_b. There are $\binom{l}{b}$ of these subsets, so the total number of inclusion counts must be $\lambda_b\binom{l}{b}$. A total of $\binom{m}{b}$ inclusion counts are generated by each block of m items, so the total number of inclusion counts must be $R\binom{m}{b}$. Equating these two expressions provides the relation

$$\binom{l}{b}\lambda_b = R\binom{m}{b}. \tag{7.2}$$

It can be shown that a uniform b-design with $b > 0$ is also a uniform $(b-1)$-design, where the relation between the inclusion counts of the subsets is given

$$\lambda_{b-1} = \lambda_b \frac{l-b+1}{m-b+1}. \tag{7.3}$$

Importantly, even if the equation (7.2) holds for a given configuration of integers (l,m,R,b), a b-design with a matching specification may still not exist.

One particularly well studied type of design is the uniform 2-design, commonly called the balanced incomplete block design (BIBD) [170]. An example of one is presented in figure 7.1. The complement of a BIBD is also a BIBD.

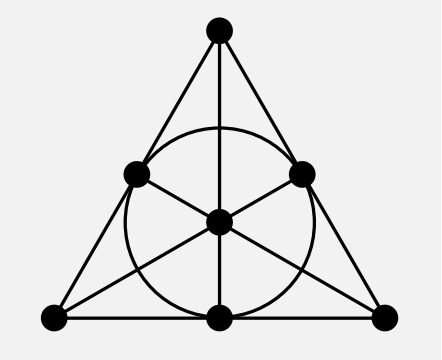

Figure 7.1: The Fano plane. This can be viewed as a BIBD with $l = R = 7$, $m = \lambda_1 = 3$, and $\lambda_2 = 1$. The circles represent the 7 indices $\{1,2,...,l\}$, the lines (including the circle) represent the subsets comprising $\mathbf{I}$, and intersection represents membership. Alternatively, the lines can be taken to represent the indices, and the circles can be taken to represent the subsets comprising $\mathbf{I}$.

7.1.2 Use in cross validation

The idea of using block designs to describe CV experiments was introduced in section 3.3.2. To reiterate, where the full sample is denoted $\mathbf{D} = \langle D_i \rangle_{i=1}^{l}$, the rth index subset $I_r = \{\imath_1, \imath_2, ..., \imath_m\}$ of a design $\mathbf{I}$ specifies the construction of the rth training set $\mathbf{G}_r = \langle G_{r,i} \rangle_{i=1}^{m}$ through the relation

$$G_{r,i} = D_{\imath_i}. \tag{7.4}$$

The testing sets are specified in the same way by the complement design **J**. All CV strategies in which the training sets comprise a fixed m items are described by uniform designs.

There is an immediate connection between designs with balanced inclusion counts and low variance CV strategies. As described in section 6.2.2, CV strategies where all items are used an equal number of times can be expected to have the lowest variance under fixed predictor models. Because equal use in testing implies equal use in training, these designs will all be 1-designs.

For a given training set sizes m, the lowest variance CV strategy is leave-p-out cross validation (LPOCV). By definition, LPOCV is described by an m-design with $\lambda_m = 1$. This design is also a b-design for $b \in \{1,2,...,m\}$.

7.1.3 Balanced incomplete cross validation

The idea of using BIBDs for CV was first proposed by Shao in 1993 [166] as an alternative to LPOCV for selecting linear models when the latter is not computationally feasible. The resulting strategy was called balanced incomplete cross validation (BICV). The precise motivation for BICV is not described, but one of Shao's key intentions may have been to enforce the equal use constraint of KCV.

In 2014, a technical report by Fuchs and Krautenbacher revisited BICV as a way to provide more efficient (i.e., lower variance for a given amount of computation) CV strategies than KCV or RKCV [171]. The same work was developed into a paper published in March of 2016 [135]. In it, the authors develop the covariance model of [167] to extend to the component train-test experiments of arbitrary CVs strategies with a fixed training set size. This model subsumes the more specific models of [37] and [38]. At the end of this development, they show that the variance of a CV performance estimate with R train-test experiments is of the form

$$\frac{1}{R^2}\sum_{i=0}^{m}\alpha_i B_i\,, \tag{7.5}$$

where the B_i represent the sum of squared inclusion counts of all subsets of $\{1,2,...,l\}$ of size i. This may be more formally expressed as

$$B_i = \sum_{\theta\in\Theta_i} a_\theta^2 \text{ in which} \tag{7.6}$$

$$\Theta_i = \{\theta \subset \{1,2,...,l\} \text{ such that } |\theta| = i\}\,. \tag{7.7}$$

The α_i in (7.5) may be positive or negative, and they are defined by

$$\alpha_i = \sum_{c=0}^{i} (-1)^{i-c} \binom{i}{c} \rho_c \,, \tag{7.8}$$

where ρ_c is the covariance between two train-test experiments in a given sample sharing c items in their training sets. This is itself defined

$$\rho_c = \mathbb{C}\mathrm{ov}\left[\gamma(u, \mathbf{G}_r, \mathbf{H}_r), \gamma(u, \mathbf{G}_{r'}, \mathbf{H}_{r'}) \right], \tag{7.9}$$

where $\mathbf{G}_r$ and $\mathbf{H}_r$ are defined by the index sets I_r and $J_r = \{1, 2, ..., l\} \setminus I_r$ in the standard way, and $I_r \cap I_{r'} = c$.

In a toy problem which is essentially equivalent to mean estimation where items are specified by the pair (X, Y) where $Y = X$, Fuchs and Krautenbacher [135] show that ρ_c is a polynomial of order 2. From the key result

$$\sum_{c=0}^{i} (-1)^{c-i} \binom{i}{c} c^j = 0 \text{ for all } j < i, \tag{7.10}$$

one can show that $\alpha_i = 0$ for all $i > 2$. Only the first three α terms contribute to the variance of a design in this case, so it may be written as follows:

$$\frac{1}{R^2} \left(\alpha_0 B_0 + \alpha_1 B_1 + \alpha_2 B_2 \right). \tag{7.11}$$

As demonstrated in the supplementary document of [135], α_1 and α_2 are strictly positive[1] To minimise this variance, one must minimise both B_1 and B_2. From its definition in equation (7.6), one can see that B_1 is minimised in any strategy where items are used an equal number times. To minimise the variance further, one must also look to minimise B_2 by ensuring that all *pairs* of items are used an equal number of times also. If a BIBD exists for a given l, m, and R, then it will minimise both B_1 and B_2. *Thus, BICV will provide a minimum variance CV strategy for a given R train-test repetitions, being more efficient than commensurate RKCV with an equivalent number of train-test repetitions.*

In the rest of their paper, Fuchs and Krautenbacher argue for BICV to be used over RKCV in more general contexts. Though no proof of variance minimisation is possible in the general case, they argue that real problems should be similar to the toy problem analysed. They include a numerical demonstration where BICV offers significant reductions in variance in a real re-

[1] To see why α_1 is positive, one must consider the final set of equation in that document, and consider that, in the notation used there, $n \geq g + 1$.

gression problem. This is encouraging, as it suggests that BICV may be used to increase the precision of performance measurement for little computation costs in small-sample applications such as AD ANA.

7.1.3.1 BICV and the validity of the fixed predictor model

As discussed in section 6.2.2, under the fixed predictor model, all equal use CV strategies should produce estimators with the same variance. In this case, the covariance between two train-test experiments is simply proportional to the overlap c of their testing sets. That is, for some constants C_0 and C_1,

$$\rho_c = C_0 + C_1 c\,. \tag{7.12}$$

As can be seen from an inspection of equations (7.8) and (7.10), this makes the coefficients α_i for all $i > 1$ equal to zero. This would make the variance of a CV experiment dependent solely on the individual item inclusion counts. The fixed predictor model is therefore unable to explain any benefit for BICV over RKCV with the same number of train-test experiments. *Thus, one should therefore expect the efficiency gain associated with BICV over RKCV to be greatest in those cases where the fixed predictor model is least accurate.* This will coincide with the cases where additional KCV repetitions offer the greatest benefit in RKCV (see section 6.2.2).

7.2 Approximately balanced cross validation

This section introduces approximately balanced cross validation (ABCV), a new CV strategy I have created to extend the benefits of BICV to a greater range of experimental settings, including use with stratification.

7.2.1 Limitations of BICV

While BICV can offer desirable low variance, the requirement that a BIBD exist for a given set of experiment parameters may be too constraining to make the strategy practical. Recall that the equation

$$R\binom{m}{2} = \lambda_2\binom{l}{2} \tag{7.13}$$

is necessary for a BIBD to exist. A result of this is that the lowest possible of R such that a BIBD is possible for a given value of (l,m) is

$$R_{\min} = \frac{\mathrm{LCM}\big[m(m-1), l(l-1)\big]}{m(m-1)}\,, \tag{7.14}$$

where LCM denotes the lowest common multiple function. For many values of the parameter pair (l,m), even $R_{\min}$ may be too large to be computationally feasible. (For instance, for $(l,m) =$

$(54, 17)$, $R_{\min} = 194616$). Even where $R_{\min}$ represents a feasible number of train-test cycles, because equation (7.13) is only necessary but not sufficient, it may be the case that no BIBD exists. Indeed, there is no general algorithm to determine whether a BIBD exists for a given (l, m, R, λ_2) satisfying equation (7.13) [135]. In many cases, it may be possible to guarantee the existence of a BIBD by discarding items from the sample (reducing l) and using a value of m close to l or 0 [135]. However, it is undesirable to have to do so, as this may lead to higher variance and bias.

Crucially, BICV is incompatible with the class stratification that is used to reduce variance and bias in classification problems [159]. If class stratification is sacrificed to allow for BICV, this may in many cases actually lead to a loss of efficiency.

When forced to abandoned desirable experimental settings, particularly stratification, one may simply prefer to revert the less efficient RKCV (or EKCV) strategies.

These limitations motivate the development of a new strategy similar to BICV that is based on designs that are only *approximately* balanced. Such designs will minimise variance *within* the constraints provided by classification stratification and limited computational resources. This should preserve much of the efficiency benefits of BICV while permitting a much greater range of experimental settings.

7.2.2 An algorithm for approximately balanced designs

The ABCV strategy that I have developed to extend the benefits of BICV to a greater number of experimental settings is based on the algorithm described in figure 7.2. While the algorithm behind the extended K-fold cross validation (EKCV) of section 6.4) greedy selected training sets to minimise B_1 of equation (7.6), the algorithm behind ABCV is based on a greedy minimisation of both B_1 and B_2. While I have described the algorithm in terms of minimising the inclusion counts of items in the *training sets*, because the complement of a BIBD is also a BIBD, the algorithm could equally be implemented through a consideration of the inclusion counts of items in the *testing sets* where this is more computationally efficient.

This method is compatible with class stratification. While, like the algorithm behind EKCV, this algorithm is guaranteed to distribute single item inclusion counts evenly (i.e. produce a 1 design) if this is possible, it is not guaranteed to find a BIBD for a given set of parameters l, m and R even if one exists.

7.3 Empirical validation

In order to assess the potential benefits of ABCV over existing CV strategies, I have conducted some preliminary empirical experiments in binary classification tasks to measure the reduction

```
set single item and pair inclusion counts tables to zero
for each train-test experiment index r do
    for each stratification group do
        add all group items to candidate list
        number of group items needed in training set is m_need
        while m_need > 0 do
            if an item has unique lowest single item inclusion count then
                select that item
            else
                select at random from items that will lead to lowest value of B_2 if
                  selected
            end
            add selected item to training set I_r
            and remove it from candidate list
            update single item and pair inclusion counts
            decrement m_need
        end
    end
    select J_r as {1,2,...,l} \ I_r
end
```

Figure 7.2: Description of the subset selection algorithm used for ABCV

in variance. I have used two simulated problems and two derived from real datasets. In both of these, I compare stratified ABCV to stratified EKCV, which represents a more flexible relaxation of the RKCV that is most commonly used currently. For all parameter settings considered here, EKCV is equivalent to RKCV and can be simply regarded as an implementation of it.

Using simple classification algorithms from 2.6.2, I produced the following selection of 6 learners to be evaluated:

1. the C45 decision tree,
2. KNN with $K = 3$,
3. KNN with $K = 5$,
4. KNN with $K = 7$,
5. Naive Bayes, and
6. Nearest Centroid.

The choice of simple algorithms (along with relatively small sample sizes) allowed me to conduct the very large number of experiments required to provide sufficient precision within a reasonable time.

7.3.1 Experiments on simulated problems

In order to measure the variance of a CV experiment in a simulated problem, one can simply preform the experiment on many independent datasets. For two synthetic classification tasks, I have generated 5000000 independent samples, and performed both ABCV and EKCV using different number of train-test experiments. This took a matter of days on single machine.

7.3.1.1 Simulation A

Simulation A was based on a simulated problem involving only 8 categorical features. Values were generated to provide some noisy signal informative of the labels; for each item, each feature took one of 4 random values with equal probability, with the exception of the first two features. The first feature always took a fixed value for the items of the first class, and the second feature always took a fixed value for those of the second class. For the nearest centroid classifier, it was necessary to define a distance metric. To do this, the squared distance between two points was defined as the number of features for which they had differing values.

For this simulation, I performed stratified ABCV and EKCV experiments that used training sets comprising 20 items of each class in samples comprising 30 items of each class. These numbers were chosen arbitrarily within the constrain on computational time. The number of component train-test cycles was either 6, 12, or 24.

7.3.1.2 Simulation B

Simulation B used a synthetic binary classification task based on homoskedastic Gaussian distributions. Where the labels Y of an item might take values of either 0 or 1, and the features X have a d dimensional multivariate Gaussian distribution conditional on Y such that $X|Y = y \sim N(\mathbf{M}(y), \Sigma)$, where

$$\Sigma := \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \qquad \text{and} \qquad M(y) := \begin{pmatrix} 2y-1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The parameter ρ may be varied between $-1/(1-d)$ and 1, resulting in different distribution shapes. Some example datasets derived from this item distribution are presented in figure 7.3.

Figure 7.3: Different samples generated in the simulated classification task with $d = 2$ using different values of the parameter ρ. Items of a given class share the same colour.

For simulation B, I used a parameter setting of $\rho = 0.2$ and $d = 2$. I performed stratified ABCV and EKCV experiments that used training sets comprising 15 items of each class in

random samples comprising 20 items of each class. The number of component train-test cycles was either 8, 16, or 24.

7.3.2 Experiments on real datasets

In order to measure the variance of a CV experiment without bias in a real dataset, one must use a resampling experiment based on disjoint subsets [35]. As an initial choice, I used two classification datasets from the KEEL collection[2]. I did this because I though a successful BICV-like strategy would be of interest to a wider machine learning audience who would expect demonstrations on standardised datasets. Had results on these been more promising, I would also have included a dataset from AD ANA. For both datasets, I drew as many disjoint subsamples as possible for a given l experiment specification, and performed both ABCV and EKCV to generate independent performance estimates which could be used to estimate variance. This division and estimation was repeated 50000 times for both datasets. This took a matter of days on single machine.

The difference in variance estimates associated with a given division/validation repetition is an unbiased estimator of the true difference in variance between ABCV and EKCV. The various difference estimates are conditional random variables; I have performed a one-sided t test to demonstrate *convergence*. The p values relate to the null hypothesis that the variance differences measurements produced by the procedure on *the given dataset* under random permutations have a mean of zero.

7.3.2.1 Banana

The banana dataset comprises 2924 items of one class and 2376 of another. This is an artificial dataset where there are two classes of items with banana shaped distributions in a feature space with two real valued features. Each division/validation iteration used 97 disjoint subsets comprising 30 items of the first class, and 24 of the second. ABCV and EKCV experiments used 32 train-test cycles with training sets comprising 15 items of the first class and 12 of the second.

7.3.2.2 Pima

The Pima dataset comprises 500 items of one class and 268 of another. This is a real dataset with seven real valued features relating medical variables that may be used to predict whether a person has diabetes. Each division/validation iteration used 8 disjoint subsets comprising 60 items of the first class, and 30 of the second. ABCV and EKCV experiments used 27 train-test cycles with training sets comprising 40 items of the first class and 20 of the second.

[2] http://sci2s.ugr.es/keel/category.php?cat=clas

7.3.3 Experimental results

Detailed results of all experiments are presented in the tables at the end of this chapter. The mean performance estimates were identical to four significant digits for both ABCV and EKCV runs, as the number of repetitions was so large. (Recall that the expectations of these methods performance estimations have identical expectations by definition.)

Figures 7.4 and 7.5 illustrate the variance results of the synthetic classification problems. In both problems, there is a clear evidence for a small reduction in variance for all learners (of the order of 1%). The reduction in variance appears to increase with the number of train-test cycles.

Results in the real datasets, presented in tables 7.3 and 7.4, show a similar outcome. ABCV appears to offer a small reduction in variance.

For the banana dataset, the p values of the convergence t test are 0 to numerical precision for all performance quantities (accuracy, specificity and sensitivity) for all learners. For the Pima dataset, the p values are all less than 0.01. I take this as an assurance of convergence in both cases.

7.4 Discussion

While ABCV did offer some reduction in variance over the RKCV now commonly used, the reduction was of negligible practical significance. This level of reduction is far shy of that achieved by BICV in the regression problems of [135], and is not great enough to justify the greater experimental complexity entailed in using ABCV over EKCV and RKCV.

This may be because even an optimally balanced design would not have offered a great reduction, or it may have been due to some limitations in the design selection algorithm. Noting that the fractional reductions in variance increased with the number of train-test cycles, it is possible that ABCV might be more useful in situations where the number of train-test cycles is of the order of 100s rather than 10s.

An optimistic reading of these results is that stratified EKCV (or RKCV) remains essentially the most efficient choice of CV strategy for classification problems where stratification is important. As discussed in 2.1, this is the dominant type of problem in AD ANA.

7.4.1 Possible improvement

It is possible that a better design selection algorithm, based on something other than greedy optimisation, might achieve greater reductions in variance. One option for extending ABCV would be to use any method of generating true or approximate BIBDs for the items of each stratification group separately, before ‘knitting’ these designs together in such a way as to min-

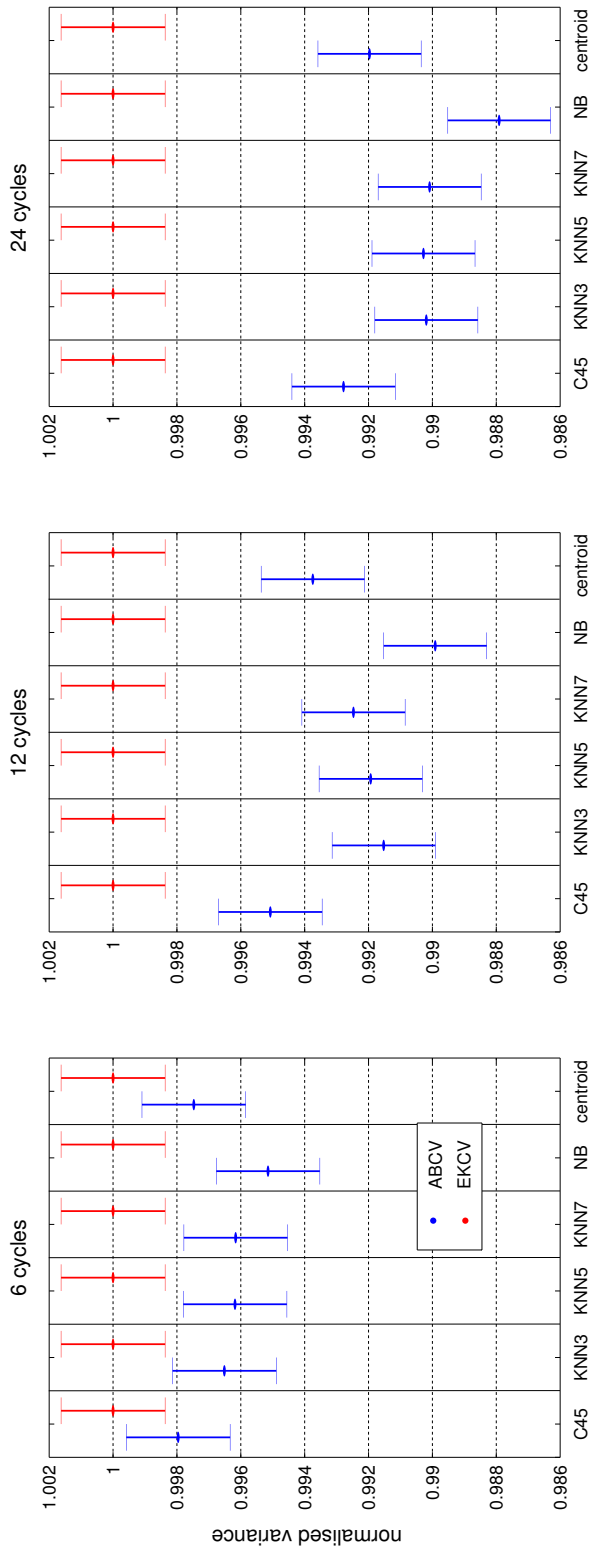

Figure 7.4: Results of experiments with simulation A. Error bars indicate 99% confidence intervals for the variances of the ABCV and EKCV methods. Variances for each learner have been normalised so that the variance associated with EKCV is 1. ABCV offers a variance reduction over EKCV that is statistically, but not practically, significant.

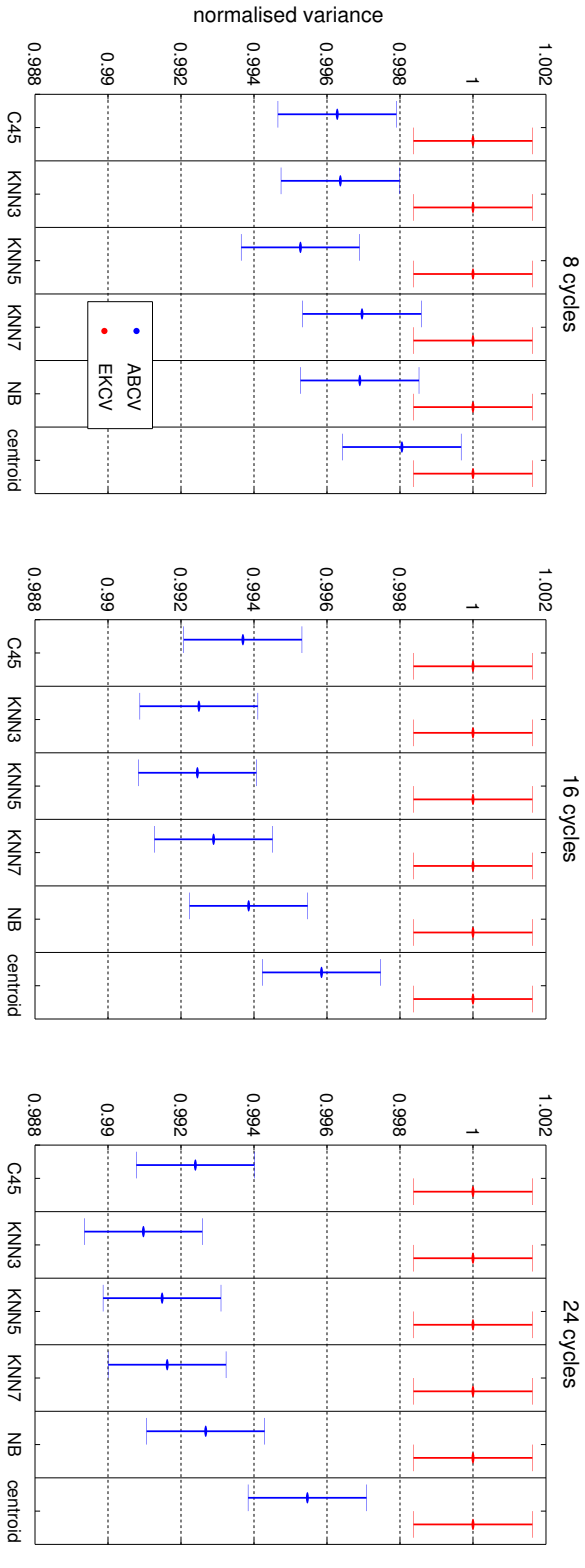

Figure 7.5: Results of experiments with simulation B. Error bars indicate 99% confidence intervals for the variances of the ABCV and EKCV methods. Variances for each learner have been normalised so that the variance associated with EKCV is 1. ABCV offers a variance reduction over EKCV that is statistically, but not practically, significant.

imise the average squared pair inclusion count. While more advanced methods of finding true or approximate BIBDs might be more computationally demanding, this problem could be solved through a repository of precomputed design elements.

Another potential extension would involve an algorithm that sought to minimise B_i for i greater than 2 in addition to B_1 and B_2. This would require a level of computational resources that increased rapidly with i, as it would require storing inclusion counts for the $\binom{l}{i}$ subsets of size i. It is not clear that it would offer a great deal of benefit, as the α_i terms at higher i values may not be large and positive.

	acc	spec	sens
C45	0.8735	0.8760	0.8710
KNN3	0.7980	0.7766	0.8193
KNN5	0.8205	0.8461	0.7949
KNN7	0.8193	0.9095	0.7291
NB	0.8667	0.8753	0.8582
centroid	0.8728	0.8729	0.8727

(a) Performance

	acc	spec	sens
C45	0.9706	3.1583	3.2774
KNN3	1.6992	2.6168	2.1404
KNN5	1.5002	1.7849	2.4949
KNN7	1.4213	1.0196	3.3952
NB	0.9136	1.1618	1.5142
centroid	0.8842	1.2333	1.2344

(b) Variance of EKCV with 6 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	0.8966	2.8356	2.9387
KNN3	1.4851	2.1970	1.7841
KNN5	1.3220	1.4912	2.1398
KNN7	1.2643	0.8417	2.9890
NB	0.7911	0.9993	1.1627
centroid	0.8023	1.0735	1.0752

(c) Variance of EKCV with 12 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	0.8601	2.6761	2.7671
KNN3	1.3790	1.9877	1.6090
KNN5	1.2332	1.3438	1.9653
KNN7	1.1868	0.7527	2.7891
NB	0.7304	0.9191	0.9876
centroid	0.7626	0.9952	0.9976

(d) Variance of EKCV with 24 cycles ($\times 10^{-3}$)

Table 7.1: Results of simulation A

	acc	spec	sens
C45	0.7906	0.7891	0.7920
KNN3	0.8034	0.8034	0.8033
KNN5	0.8145	0.8146	0.8144
KNN7	0.8198	0.8199	0.8196
NB	0.8229	0.8230	0.8227
centroid	0.8324	0.8325	0.8324

(a) Performance

	acc	spec	sens
C45	3.3211	4.4207	4.5907
KNN3	2.8571	3.8980	3.9267
KNN5	2.4486	3.7907	3.8229
KNN7	2.1941	3.7847	3.8096
NB	1.9274	3.9348	3.9667
centroid	1.6537	2.5249	2.5206

(b) Variance of EKCV with 8 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	3.1192	3.9424	4.1137
KNN3	2.7060	3.6107	3.6382
KNN5	2.3184	3.5496	3.5742
KNN7	2.0803	3.5725	3.5906
NB	1.8360	3.7659	3.7890
centroid	1.5983	2.4179	2.4110

(c) Variance of EKCV with 16 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	3.0546	3.7845	3.9563
KNN3	2.6574	3.5181	3.5424
KNN5	2.2755	3.4694	3.4952
KNN7	2.0423	3.5013	3.5204
NB	1.8063	3.7080	3.7336
centroid	1.5796	2.3826	2.3748

(d) Variance of EKCV with 24 cycles ($\times 10^{-3}$)

Table 7.2: Results of simulation B

	acc	spec	sens
C45	0.6833	0.7632	0.5235
KNN3	0.6876	0.8106	0.4417
KNN5	0.6947	0.8426	0.3990
KNN7	0.6968	0.8645	0.3613
NB	0.7225	0.8649	0.4377
centroid	0.6426	0.7175	0.4930

(a) Performance

	acc	spec	sens
C45	2.1799	1.9484	4.4909
KNN3	2.2036	1.8410	8.4410
KNN5	1.9456	1.7560	10.1952
KNN7	1.7237	1.7590	11.7576
NB	1.7096	3.4175	12.3356
centroid	2.1800	4.2003	4.7905

(b) Variance of EKCV with 27 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	0.9966	0.9913	0.9887
KNN3	0.9936	0.9907	0.9885
KNN5	0.9931	0.9924	0.9918
KNN7	0.9921	0.9942	0.9934
NB	0.9934	0.9975	0.9954
centroid	0.9898	0.9942	0.9825

(c) Ratio of ABCV variance to EKCV variance

Table 7.3: Results on Pima dataset

	acc	spec	sens
C45	0.6617	0.7095	0.6019
KNN3	0.7486	0.8281	0.6493
KNN5	0.7058	0.7967	0.5922
KNN7	0.6549	0.7412	0.5470
NB	0.5596	0.6553	0.4399
centroid	0.5102	0.5127	0.5072

(a) Performance

	acc	spec	sens
C45	3.4010	3.2542	5.7070
KNN3	2.5919	2.5642	6.3962
KNN5	2.8686	3.0336	9.3353
KNN7	3.3426	4.1610	12.8570
NB	3.5567	12.6525	25.6415
centroid	4.0695	4.3040	5.6326

(b) Variance of EKCV with 32 cycles ($\times 10^{-3}$)

	acc	spec	sens
C45	0.9873	0.9652	0.9710
KNN3	0.9776	0.9564	0.9754
KNN5	0.9752	0.9448	0.9834
KNN7	0.9750	0.9437	0.9884
NB	0.9850	0.9886	0.9950
centroid	0.9864	0.9691	0.9580

(c) Ratio of ABCV variance to EKCV variance

Table 7.4: Results on Banana dataset

Part IV

Statistical procedures

Chapter 8

Statistical inference in performance estimation

In this chapter, I shall introduce the key concepts required to understand frequentist statistical analysis in the context of learner and predictor performance estimation based on CV results. This will provide the necessary theoretical background for the discussion of specialist statistical procedures in chapters 9 and 10. After an abstract description of the frequentist inference, I shall present the normal and binomial models that one may use for inference for the performance of a fixed predictor in a simple testing experiment. I shall then introduce the problem of dependency, which occurs when the distributions of test statistics deviate from those expected under the model used for inference. I shall describe how the distributions of performance results in SHOCV and various other CV experiments deviate from those expected under fixed predictor models, and what this means for inference strategies based on those models. Finally, I shall describe the concepts of repeatability and replicability in statistical testing.

Importantly, this thesis is concerned with inference for a learner performance in a specific learning problem. I do not consider inference for the average performance of learners across families of problems. A thorough treatment of that subject is found in [172].

8.1 Introducing frequentist inference

Statistical inference is concerned with deducing the parameters of some random variable's underlying distribution on the basis of an observed sample of i.i.d. observations. The deductions it produces come in the form of statements about the unknown parameters that have statistical correctness guarantees. The dominant approach to statistical inference in the biological and medical sciences is frequentism, in which the statements are guaranteed to be correct in a given proportion of independent experiment repetitions (independent sequences of observations).

This thesis is primarily concerned with the assessment of predictors and/or learners in a particular learning problem (rather than across families of learning problems). In this task, the observations take the form of one or more sequences $\langle Q_i \rangle_{1 \leq i \leq n}$ of identically distributed

performance measurements describing the utility of the predictions made on a set of n labelled items in a testing set $\mathbf{H} = \langle H_i \rangle_{1 \leq i \leq n}$. As described in chapter 3, where a predictor t is being studied, the performance on an item $H_i = (X_i, Y_i)$ is given $Q_i = \phi\big(t(X_i), Y_i\big)$. When considering a single sequence of predictions (from a single predictor or learner), the parameter of interest is normally the mean $\mu_t = \mathbb{E}[Q_i]$, which describes the expected utility of future predictions. When two sequences of predictions on the same set of items are being compared, these may be denoted $\langle {Q^{(1)}}_i \rangle_{1 \leq i \leq n}$ and $\langle {Q^{(2)}}_i \rangle_{1 \leq i \leq n}$. The parameter of interest in this case is normally the difference between the means of the two $\mu_d = \mu_1 - \mu_2 = \mathbb{E}\big[Q_i^{(1)} - Q_i^{(2)}\big]$, which reflects the relative utility of the two prediction types.

There are two main types of frequentist analysis: hypothesis testing and confidence interval estimation. Both of these rely on some test statistic Ξ derived from the Q_i. This has a known distribution for a given value of the fixed but unknown mean parameter μ_t. Let η denote a possible value of μ_t, and ξ denote a possible value of Ξ. For all plausible values of ξ and η, one must have a model to produce

$$P(\Xi \leq \xi \,|\, \mu_t = \eta)\,. \tag{8.1}$$

8.1.1 Hypothesis testing

Hypothesis testing relies on some univariate test statistic Ξ for which smaller values become more likely when μ_t is smaller, and larger values become more likely when μ_t is larger. Before any experiment is carried out, a **null hypothesis** H_0 of the form $\mu_t \leq \eta$, $\mu_t \geq \eta$, or $\mu_t = \eta$ is formulated. This null hypothesis should reflect some default assumption against which sufficient evidence must be amassed before it can be discounted. Its negation is the alternative hypothesis H_1.

In the case where H_0 is of the form $\mu_t \geq \eta$, an observation $\Xi = \xi$ produces a value p specified

$$p(\xi) = P(\Xi' \leq \xi \,|\, \mu_t = \eta)\,, \tag{8.2}$$

where Ξ' represents an independent observation of the random variable Ξ. This statistic is termed a p**-value**. In the case where $\mu_t = \eta$, the probability that $p(\Xi) \leq \alpha$ is less than or equal to α. If $\mu_t > \eta$, the probability that $p(\Xi) \leq \alpha$ is even less. Where H_0 is of the form $\mu_t \leq \eta$ rather than $\mu_t \geq \eta$, p is instead defined $p(\xi) = P(\Xi' \geq \xi \,|\, \mu_t = \eta)$. When H_0 is of the form

$\mu_t = 0$, p may be defined

$$p(\xi) = 2\min\Big(p_l(\xi), p_u(\xi)\Big), \text{ where}$$
$$p_l(\xi) = P(\Xi' \leq \xi \,|\, \mu_t = \eta) \text{ and}$$
$$p_u(\xi) = P(\Xi' \geq \xi \,|\, \mu_t = \eta).$$

Under all types of H_0, the definitions of p ensure that $P(p \leq \alpha \,|\, H_0) \leq \alpha$. p values may be used to express the degree of evidence against H_0, with smaller values representing greater evidence. Procedures in which H_0 is of the form $\mu_t \geq \eta$ or $\mu_t \leq \eta$ are called **one-sided**; procedures in which H_0 is of the form $\mu_t = \eta$ are called **two-sided**. A null hypothesis **significance test** (NHST) is conducted by selecting some H_0 and some value of α, termed the **significance level**, before an experiment is conducted and value of Ξ is observed. Based on the observed value of Ξ and the induced value of p, one of the following decisions is made:

- if $p \leq \alpha$, there is significant evidence against H_0, and it should be *rejected* in favour of H_1; or
- if $p > \alpha$, there is not significant evidence against H_0, and one should *draw no conclusions until the arrival of further evidence*.

This procedure can produce two types of error: **type I** errors, in which the H_0 is falsely rejected when it is true, and **type II** errors, in which H_0 is not rejected when it is false. Because

$$P(p \leq \alpha \,|\, H_0) \leq \alpha,$$

the probability of a type I error is strictly limited by α. This is the correctness guarantee offered by NHST. In the medical and biological sciences, the most common choice of α is 0.05. The decision threshold α used to make a decision based on the observed value of p induces a decision threshold ξ_α which may be used to make the same decision based on the observed value of Ξ.

When Ξ is a continuous variable with a finite probability density at all points, then p will be uniformly distributed over the interval $[0,1]$, and the type I error of a test will be precisely α. Where Ξ is discrete, p will also be discrete, and the true type I error rate may be less than the nominal rate of α. In practice, an inexact model for Ξ may also cause the type I error deviate from the nominal α. A test in which the type I error is less than the nominal value is termed **conservative**. In this thesis, a test in which the type I error is greater than the nominal value

shall be called **permissive**, although the terms anti-conservative, liberal, and invalid are also used elsewhere.

While the type I error rate may not be greater than a specified value, there is no such limit of the type II error rate. The power of a test is the probability with which it rejects the null hypothesis when the null hypothesis is not true. (Equivalently, power can be defined as one minus the type II error rate.) Power is therefore dependent on the true value of μ_t and the degree of deviation from the null hypothesis that that entails. Crucially, one never 'confirms' the null hypothesis in NHST based on $p > \alpha$; because there is no lower limit on the power, there is no guarantee on the probability that such a statement would be correct.[1]

An illustration of the distribution of a significance test against the null hypothesis $\mu_t \geq \eta$ is provided in figure 8.1.

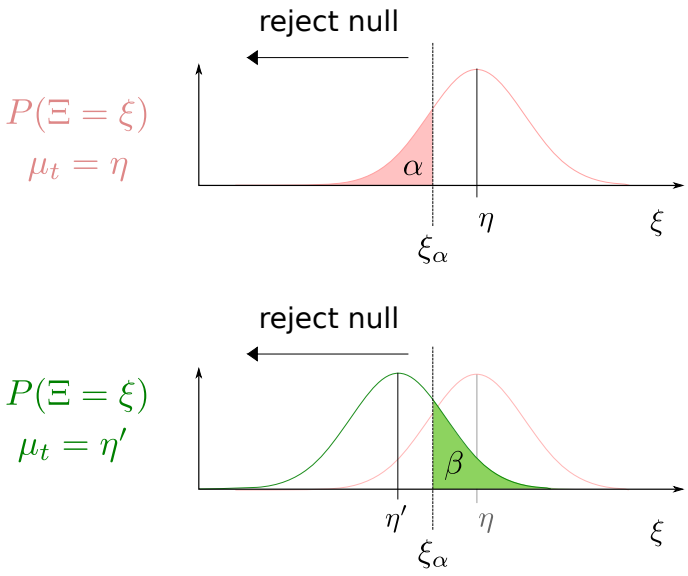

Figure 8.1: Illustration of a significance test. Above, the distribution of Ξ under the null hypothesis is used to define a threshold ξ_α that limits the probability of type I errors to α. Below, the distribution of Ξ in a non-null situation determines a type II error rate β for the same test when μ_t takes a particular value η'.

8.1.2 Confidence intervals

In NHST, one begins with a pre-specified null hypothesis, and the observed value of a test statistic Ξ determines the level of evidence against it as represented by the p-value. This is used to make a decision based on a pre-determined significance level α. In the construction of

[1] There is an analogy with the presumption of innocence in law: a person accused of a crime is considered innocent (H_0) until the possibility is ruled out beyond reasonable doubt ($p < \alpha$). Even though a court may believe it is more likely that the accused is guilty than innocent, there may not be sufficient evidence to convict. In this case, it is not that the accused has been 'found innocent'. Rather, the default decision is made due to lack of sufficient evidence.

confidence intervals, the observed value of a test statistic Ξ is used to select a hypothesis such that the level of evidence against it reaches a given pre-determined value α.

To place an upper bound on μ_t after a value of Ξ has been observed, one needs a function b_u defined

$$b_u(\xi) = \arg\max_{\eta} P(\Xi \leq \xi \,|\, \mu_t = \eta) \text{ such that } P(\Xi \leq \xi \,|\, \mu_t = \eta) \leq \alpha\,. \tag{8.3}$$

The restriction in the above definition ensures that, for all ξ,

$$P\big(\Xi \leq \xi \,|\, \mu_t = b_u(\xi)\big) \leq \alpha\,. \tag{8.4}$$

The function b_u must be monotonic non-decreasing with ξ. This ensures that, for all ξ,

$$\begin{aligned} P\big(\Xi \leq \xi \,|\, \mu_t = b_u(\xi)\big) &= P\big(b_u(\Xi) \leq \mu_t \,|\, \mu_t = b_u(\xi)\big) \\ &\leq \alpha \text{ by (8.4).} \end{aligned} \tag{8.5}$$

Because this does not depend on the value of ξ, one can simply write

$$P\big(b_u(\Xi) \leq \mu_t\big) \leq \alpha\,. \tag{8.6}$$

Thus,

$$\begin{aligned} P\big(\mu_t < b_u(\Xi)\big) &= 1 - P\big(b_u(\Xi) \leq \mu_t\big) \\ &> 1 - \alpha\,. \end{aligned} \tag{8.7}$$

The statement $\mu_t \in \big(-\infty, b_u(\Xi)\big]$ is therefore true with a probability of at least $1-\alpha$. For this reason, $\big(-\infty, b_u(\Xi)\big]$ may be called a $(1-\alpha)$ one-sided **confidence interval** for μ_t. Any null hypothesis about μ_t inconsistent with μ_t's membership of the interval may be rejected. While there is a guaranteed rate at which the true value of μ_t is contained, there is no guaranteed rate at which incorrect values of μ_t are excluded. By analogy with NHST, cases where the interval does not contain the true value of μ_t may be termed type I errors. The rate at which an interval contains the true value of μ_t is termed the **coverage**. Interval procedures with lower than nominal coverage may be called permissive; procedures with greater than nominal coverage may be termed conservative.

A lower bound on μ_t may be produced in a similar way to the upper bound just described, with the end result being a procedure in which $P(\mu_t > b_l(\Xi)) > 1-\alpha$. To build a two-sided

interval with lower and upper bounds b_l and b_u, one selects these in the same way as the one-sided intervals, but with half the value of α. This means that

$$P\big(\mu_t < b_u(\Xi)\big) > 1-\alpha/2\,, \text{ and}$$
$$P\big(\mu_t > b_l(\Xi)\big) > 1-\alpha/2\,.$$

Thus, providing $b_u(\Xi) \geq b_l(\Xi)$,

$$P\big(b_l(\Xi) < \mu_t < b_u(\Xi)\big) > 1-\alpha\,. \tag{8.8}$$

A one-sided interval with $P\big(\mu_t < b_u(\Xi)\big) > 1-\alpha$ may be presented as an interval with two bounds b_l and b_u, where b_u is defined as before and $b_l = -\inf$.

An illustration of confidence interval construction is presented in figure 8.2.

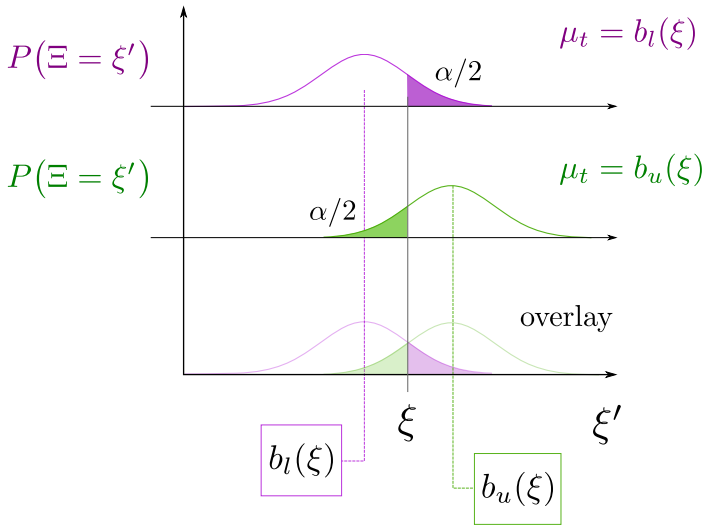

Figure 8.2: Illustration of a $(1-\alpha)$ confidence interval construction. When a value of $\Xi = \xi$ is observed, bounds b_l and b_u are selected such the probability of obtaining a value of Ξ any more extreme from the mean is no more than $\alpha/2$. This ensures that for any true value of μ_t, the probability of observing a value of Ξ sufficiently extreme to exclude μ_t from a confidence interval is less than α.

8.1.2.1 Approximate intervals

In cases where μ_t and Ξ are continuous, and Ξ has a distribution with finite density at all points, intervals constructed according to equation (8.8) will have a coverage precisely equal to the nominal value of $(1-\alpha)$. Where Ξ is discrete or μ_t may take only one of a discrete set of values, these intervals will tend to be conservative.

Sometimes, in cases where the exact interval construction procedure outlined in this sec-

tion produces a very conservative procedure, one may prefer to use an approximate interval procedure. Such a procedure will produce intervals such that

$$P\big(b_l(\Xi) < \mu_t < b_u(\Xi)\big) \approx 1 - \alpha \tag{8.9}$$

for all values of μ_t. While coverage may be slightly below the nominal value in some problems (for some possible values of μ_t and other unknown population parameters), it may be closer to it than the higher coverage produced by the conservative alternative. A small drop in coverage may be an acceptable sacrifice when it produces much narrower intervals.

8.1.3 Contrast with Bayesian inference

In frequentist inference, the unknown population parameters are taken as fixed, and the statements about them are guaranteed to be correct in a given fraction of independent experiment repetitions sharing a fixed value of μ_t. Frequentist inference may be contrasted with Bayesian inference, in which the population parameters are considered to be the result of a random process, and are treated as random variables generated according to some **prior** distribution. After a value of the test statistic value $\Xi = \xi$ is observed, this induces a posterior distribution for the population parameters according to Bayes' rule. In the case where the sole parameter of interest is μ_t, and the sole statistic is Ξ, the posterior is specified

$$P(\mu_t = \eta | \Xi = \xi) = \frac{P(\Xi = \xi | \mu_t = \eta) P(\mu_t = \eta)}{P(\Xi = \xi)}, \tag{8.10}$$

where $P(\mu_t = \eta)$ represents the prior distribution of μ_t. This posterior may be used to assign probabilities to hypotheses about the value of μ_t conditional on the observed value of $\Xi = \xi$. These probabilities may be interpreted as describing the fraction of experiments in which μ_t lies in the interval after that value of $\Xi = \xi$ has been observed. Rather than μ_t taking a fixed, shared value across all the hypothetical experiments, each experiment has a randomly generated μ_t according to the prior distribution.

The prior distribution is often taken to represent one's subjective beliefs about the distribution of the parameters in problems that appear like the one under study. This may differ between researchers, meaning that different experiments may draw different conclusions about the relative probability of various hypotheses about μ_t based on the same observations. Accordingly, a Bayesian analysis may not convince others to share a researcher's conclusion if they do not share the researcher's prior. This limitation, along with a tendency toward tradition, may be the reason that frequentist analysis methods are still dominant in many areas of natural and clinical

science[2]. This, in turn, is the reason for the exclusive focus on frequentist methods in this part of the thesis. Note that Bayesian analysis still requires one to know the distribution of the test statistic given the unknown population parameters, so it will still suffer from the problem of dependency.

8.1.4 The role of inference in ANA

The methods of ANA research have intended applications in pre-selection and response tracking in clinical trials, or in clinical decision making. In both of these contexts, the introduction of ANA methods will modify an existing procedure to produce results that are better or worse. If the introduction of ANA methods actually leads to poorer results, then there is the potential for real harm; poorer clinical decision harm patients directly, and inefficient clinical trials waste resources and may expose subjects to unnecessary risks. Statistical analysis can be used to guarantee that new methods offer a genuine performance improvement and so limit the potential for harm.

Confidence intervals in particular are important for ensuring the generalisation of results in small-sample supervised learning applications [173]. Without them, it is very difficult to assess the amount of information provided by a study, and it is difficult to interpret results across multiple studies. Figure 8.3 depicts an imagined meta-analysis considering the results of multiple studies. Without confidence intervals or some other measure of uncertainty, there is no way for an interpreter to efficiently combine the results across the various studies, to know the accuracy of the combined estimate, or to identify when different studies are likely to be in disagreement.

Pairwise significance testing can be used to establish that one learner has greater performance than another. This is good if one wishes to guard against false conclusions that a change in learner offers improved performance, but it is not useful if one's goal is simply the identification of the best learner. In the latter case, there is no need to require statistical significance to believe one learner is likely to have a greater performance than another. As discussed in section 8.1.1, significance testing requires some default decision to be made unless the evidence to the contrary reaches a required threshold. This is appropriate in clinical trials when there is an established treatment with well known characteristics over which a new therapy must demonstrate benefit. However, it is not appropriate when deciding which of two learners, *about which a comparable amount is known*, is superior. Significance testing is also no substitute for confidence intervals, as it cannot be used to interpret results over multiple studies [174]. Of course, the two analyses are not exclusive.

[2] though a Bayesian interpreter might also reject the conclusion of a frequentist analysis

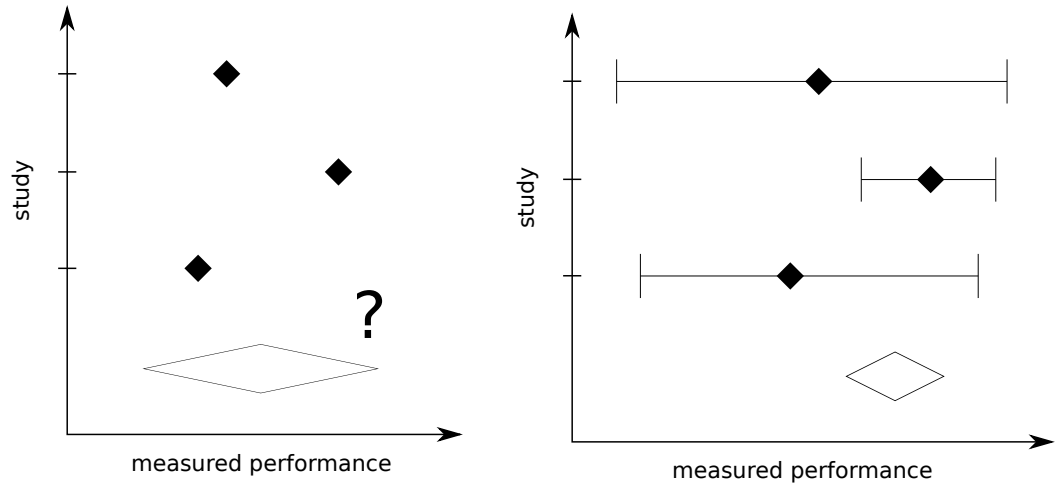

Figure 8.3: Intervals in the interpretation of multiple studies. Left: without some reporting of intervals or measurement uncertainty, it is difficult to combine measurements of independent studies of a single quantity. Right: with intervals, it is possible to combine the various measurement to produce an efficient estimator for the quantity with well known uncertainty.

8.2 Fixed predictor models for performance inference

In this section, I shall introduce two statistical models used for significance testing and interval construction in performance estimation tasks. The first of these is the normal model, which is very generally applicable, and the second is the binomial model, which is applicable to classification tasks. Both models assume that the individual observations in a series $\langle Q_i \rangle_{1 \leq i \leq n}$ of performance measurements are i.i.d., and use that fact to infer their shared mean μ_t. As such, they are all appropriate in the case where these measurements have been produced by a fixed predictor t on a testing set of i.i.d. items.

8.2.1 Normal models

Under the normal model, the Q_i are i.i.d. normal variables with some mean μ_t and variance σ_t^2. They may be modelled

$$Q_i = \mu_t + \sigma_t \varepsilon_i \tag{8.11}$$

where the ε_i are i.i.d. standard normal variables. The variance σ_t^2 may be either known or unknown, and this will affect how inference proceeds. In real performance estimation problems, it is normally unknown.

When the normal model is used to assess a fixed predictor, the Q_i will be taken to represent the performance measure $\phi(t(X_i), Y_i))$ on an item specified (X_i, Y_i). The mean then represents the predictor performance. When the normal model is used to compare two predictors with associated performance measure sequences $\langle {Q^{(1)}}_i \rangle_{1 \leq i \leq n}$ and $\langle {Q^{(2)}}_i \rangle_{1 \leq i \leq n}$, the Q_i will be taken

to represent the differences between those performance measures. That is,

$$Q_i = Q_i^{(1)} - Q_i^{(2)}\,. \tag{8.12}$$

In this case, μ_t represents the difference between the two predictor performances. The model is the same in both cases, and inference proceeds identically.

Where the variance σ_t^2 is known, the sample mean

$$\bar{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i\,, \tag{8.13}$$

carries all necessary information to make statements about μ_t. It is a normally distributed with mean μ_t and variance σ_t^2/n. The normalised Z-statistic $(\bar{Q}-\mu_t)/\sigma_t$ has a standard normal distribution.

When the variance is unknown, both the sample mean and the sample variance, defined

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Q_i-\bar{Q})^2\,, \tag{8.14}$$

are required. When considering a null hypothesis involving a potential μ_t value of η, these are combined to produce a t-statistic

$$\frac{(\bar{Q}-\eta)\sqrt{n}}{S}\,. \tag{8.15}$$

This has a t-distribution with $n-1$ degrees of freedom (DOF) when $\mu_t = \eta$.

8.2.1.1 Hypothesis tests

In the case where σ_t^2 is known, a p value to test against the null hypothesis $\mu_t \geq \eta$ is given by

$$p_l = \Phi^{-1}\left(\frac{(\bar{Q}-\eta)\sqrt{n}}{\sigma_t}\right), \tag{8.16}$$

where Φ is the inverse standard normal cumulative distribution function. The resulting procedure is called a **Z-test**.

In the case where σ_t^2 is unknown, a p value to test against the null hypothesis $\mu_t \geq \eta$ is given by

$$p_l = T_{n-1}^{-1}\left(\frac{(\bar{Q}-\eta)\sqrt{n}}{S}\right), \tag{8.17}$$

where T_{n-1}^{-1} is the inverse cumulative t-distribution with $n-1$ DOF. A test built on this p is termed a **t-test**.

To test against a null hypothesis $\mu_t \leq \eta$ instead, one will instead use the p-value

$p_u = 1 - p_l$. To perform a two-sided test against the null hypothesis $\mu_t = \eta$, one will use the $p = 2\min(p_l, p_u)$.

8.2.1.2 Confidence intervals

Where the variance is known, let $z_\alpha = \Phi^{-1}(\alpha)$. Because $(\bar{Q} - \mu_t)\sqrt{n}/\sigma_t$ is a standard normal variable,

$$P\Big(\sqrt{n} \cdot \frac{\bar{Q} - \mu_t}{\sigma_t} \leq z_\alpha\Big) = \alpha\,. \tag{8.18}$$

Rearranging this, and using the symmetry relation $z_\alpha = -z_{\alpha-1}$, produces

$$P\Big(\bar{Q} + \frac{\sigma_t z_{1-\alpha}}{\sqrt{n}} \leq \mu_t\Big) = \alpha\,. \tag{8.19}$$

This means that an upper bound in a $(1-\alpha)$ one-sided confidence interval may therefore be defined

$$b_u(\bar{Q}) = \bar{Q} + \frac{z_\alpha \sigma_t}{\sqrt{n}}\,. \tag{8.20}$$

Where the variance is unknown, let $\tau_\alpha = T^{-1}(\alpha)$. Because $(\bar{Q} - \mu_t)\sqrt{n}/S$ has a t distribution

$$P\Big(\frac{(\bar{Q} - \mu_t)\sqrt{n}}{S} \leq \tau_\alpha\Big) = \alpha\,. \tag{8.21}$$

Rearranging this, and using the symmetry relation $\tau_\alpha = -\tau_{\alpha-1}$, produces

$$P\Big(\bar{Q} + \frac{\tau_{1-\alpha} S}{\sqrt{n}} \leq \mu_t\Big) = \alpha\,. \tag{8.22}$$

This allows

$$b_u(\bar{Q}, S^2) = \bar{Q} + \frac{\tau_\alpha S}{\sqrt{n}}\,. \tag{8.23}$$

to define the one-sided upper bound in a $(1-\alpha)$ confidence interval.

In both cases, an upper bound may be produced by substituting $\tau_{1-\alpha}$ or $z_{1-\alpha}$ for τ_α or z_α, and a two-sided $(1-\alpha)$ interval may be produced using the upper and lower bounds from one-sided $(1-\alpha/2)$ intervals.

8.2.1.3 Generality

There are few cases where it reasonable to expect the Q_i to be exactly normally distributed. Wherever the Q_i are bounded (e.g., square error, accuracy), this cannot be the case. Nevertheless, the normal model can be usefully applied in many practical cases; even where the Q_i are not themselves normal, their mean $\bar{Q}$ will be approximately normally distributed whenever n is

large.[3] The distributions of both the Z and t statistics will converge to the normal distribution as n increases [175, chapter 14]. The maximum absolute difference between the true and modelled cumulative distributions of either statistic has a well defined limit [176]. Thus, the deviation of the true type I error rate from the nominal value is also limited.

8.2.2 Binomial Models

In classification tasks where the accuracy metric is used to assess the performance of a fixed predictor, the Q_i may only take the values 0 and 1 (corresponding to incorrect and correct predictions on item i). This makes them Bernoulli random variables whose rate parameter $\mu_t = P(Q_i = 1) = \mathbb{E}[Q_i]$ is the predictor performance. The quantity $n\bar{Q}$, in which

$$`\bar{Q} = \sum_{i=1}^{n} Q_i'$$

as before, is binomially distributed. It has a cumulative distribution given

$$P(n\bar{Q} \leq q) = B_{1-\mu_t}(n-q, q+1)\,, \tag{8.24}$$

where $B_{1-\mu_t}(n-q,q+1)$ is the regularised incomplete beta function. This is only defined for integer values of q between 0 and n inclusive. As such, there is no exact inverse cumulative distribution function, and so confidence intervals and hypothesis tests may not be exact.

8.2.2.1 Hypothesis tests

In classification tasks, one may wish to test against the null hypothesis that a fixed predictor has a performance μ_t that is no better than a chance rate η. A p value for a one-sided test against $\mu_t \leq \eta$ may be produced as follows after an observation of $\bar{Q} = q$:

$$\begin{aligned} p(q) &= P(n\bar{Q}' \geq nq) \\ &= 1 - B_{1-\mu_t}(n-nq-1, n\bar{Q}+1) \\ &= B_{\mu_t}(nq, n-nq+1) \end{aligned} \tag{8.25}$$

The associated test may be called a binomial test. Because the distribution of $n\bar{Q}$ is discrete, no value of q may exist such that $P(n\bar{Q} \geq q)$ is precisely α. A p value of less than α will occur only at or above some slightly higher q at which $P(n\bar{Q} \geq q) < \alpha$. This will make the test conservative. The difference between the nominal and actual type I error rates will tend to be higher when n is low.

[3] It is assumed that the Q_i have a well defined second moment, or inference for μ_t would not be possible in any case.

When comparing two classification predictors, say t_1 and t_2, one must consider both the number of times that each predictor was correct and the numbers of times that both predictors agreed. Let $\langle {Q^{(1)}}_i \rangle_{1 \leq i \leq l}$ and $\langle {Q^{(2)}}_i \rangle_{1 \leq i \leq l}$ denote the sequences of performance measurements associated with predictors t_1 and t_2 respectively. The results on n testing items may be summarised by the 2×2 **contingency table** below. This defines the key statistics

	t_1 incorrect	t_1 correct
t_2 incorrect	A_0	A_1
t_2 correct	A_2	A_2

$$A_1 = \sum_i Q_i^{(1)}(1 - Q_i^{(2)}),$$
$$A_2 = \sum_i Q_i^{(2)}(1 - Q_i^{(1)}), \text{ and}$$
$$A = A_1 + A_2.$$

Here, A_1 and A_2 represent the number of test items on which each predictor provided a superior prediction to the other, while A represents the number of test items in which their predictions disagreed. All of these quantities have marginal binomial distributions. When two predictors have equal performance (i.e. $\mu_1 = \mu_2$ and thus $\mu_d = 0$), each predictor is equally likely to be correct in any case in which the two disagree. This means that, conditional upon $A = a$, A_1 has a binomial distribution with a trials and a rate parameter of 0.5. To test against the null hypothesis $\mu_1 \geq \mu_2$ after observations $A = a$ and $A_1 = a_1$, one may consider the p value

$$\begin{aligned} p(a, a_1) &= P(A_1' \leq a_1 \,|\, A' = a) \\ &= B_{1/2}(a - a_1, a_1 + 1). \end{aligned} \tag{8.26}$$

For any possible value of a, $P(p(A,A_1) \leq \alpha \,|\, A = a) \leq \alpha$. Thus, $P(p(A,A_1) \leq \alpha) \leq \alpha$ marginally also, allowing p to be used as the basis of a conservative test against $\mu_1 \geq \mu_2$. The test for a difference in performance between two predictors based on this specification of p is called **McNemar's test**.

8.2.2.2 Confidence intervals

The discrete distribution of a binomial variable such as $n\bar{Q}$ makes it impossible to produce confidence intervals with a precise coverage of $(1 - \alpha)$ for a given true value of the rate parameter μ_t. For any given interval procedure, the coverage may vary appreciably with the true value of μ_t. A variety of interval methods exist that approach this problem in different ways [177]. These include the traditional or Wald method, which uses a normal approximation of the bino-

mial distribution with the same mean and variance. Unfortunately, this can lead to coverages much below the nominal values [177] when n is small or μ_t is near 0 or 1. At the cost of wider intervals, one may use the conservative Clopper-Pearson method as an alternative. This uses the true cumulative distribution of the binomial distribution to produce conservative intervals that will always have coverages at or above the nominal values. This can be very conservative for some values of μ_t. In the general case, one may prefer a compromise solution that will have coverage at or close to the nominal value for all values of μ_t while producing narrower intervals than the Clopper-Pearson method. In this thesis, I shall use the Agresti-Coull method, which is one of those recommended by Brown et al. [177]. This method can be seen as a modification of the Wald method that includes a varying additional of pseudo-observations of both binary outcomes before proceeding as before. I selected it because it has acceptable coverages and is straightforward to implement.

Coverages of the intervals methods discussed are presented in figure 8.4.

8.3 The problem of dependency

The fixed predictor models can provide reliable inference for the predictor performance based on the results of a testing experiment for a fixed predictor t. There, the fact that item performance measures Q_i are i.i.d. is used to make reliable statements about their marginal mean, the predictor performance μ_t. Fixed predictor models may also be applied to provide inference for the learner performance on the basis of the results observed in a CV experiment where a learner u produces a series of predictors $\langle T_r \rangle_{r=1}^{R}$. To use them in this case is to effectively assume that, for all r, $T_r = t$ such that $\mu_t = \mu_u$. From here on, this event is denoted $fixed$. This assumption means that they fail to account for the relationships between component item and test set performance results described in sections 8.4 and 8.5. These relationships, or **dependencies**, will cause the distributions of relevant test statistics to deviate from the those expected under the fixed predictor models. This section will describe how this in turn undermines inference strategies based on them (that is, how it causes the problem).

As described in sections 8.2.1 and 8.2.2, performance inference typically uses a univariate test statistic Ξ derived from the final performance estimate $\bar{Q}$. Positive correlations between the various $Q_{r,j}$ cause the distribution of $\bar{Q}$ to be wider about its mean, where wider may be defined

$$P(\bar{Q} \leq q) > P\big(\bar{Q} \leq q \,|\, fixed\big) \text{ for } q < \mu_u, \text{ and}$$
$$P(\bar{Q} \geq q) < P\big(\bar{Q} \geq q \,|\, fixed\big) \text{ for } q > \mu_u.$$

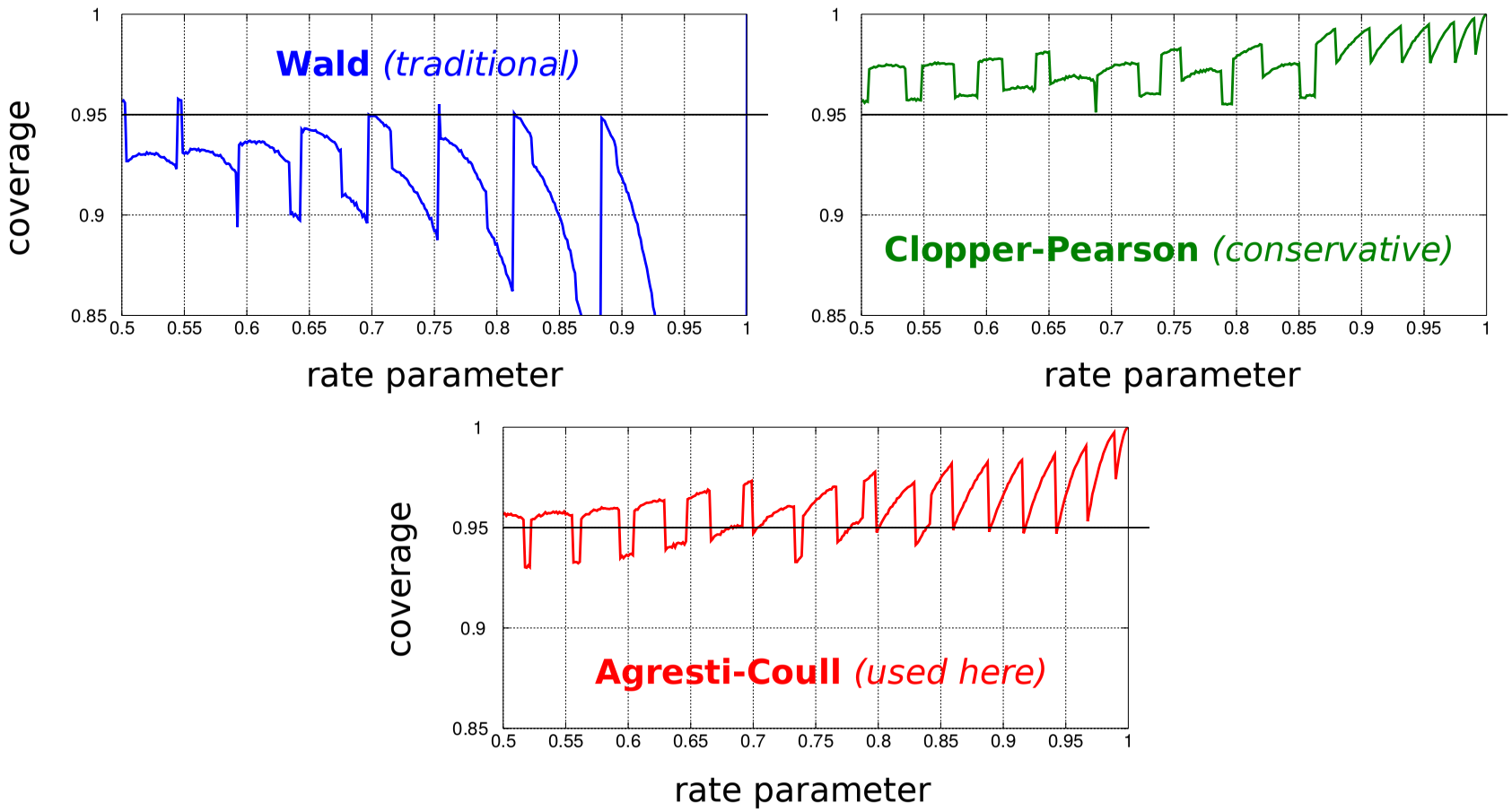

Figure 8.4: Coverages of three different 95% intervals described in section 8.2.2.2 for the rate parameter of the binomial distribution. These were calculated for using 50000 pseudorandomly generates samples of 25 observations for a grid of 5001 rates spaced equally between 0.5 and 1.0 inclusive. The systematic deviations from the nominal 95% coverage can be clearly seen.

Significance tests based on fixed predictor models reject the null hypothesis H_0 based on an observation $\Xi < \xi_\alpha$, where ξ_α has been selected to ensure this happens with a probability of no more than α in the case where this leads to a type I error (i.e., when H_0 is true). As illustrated in figure 8.5, when the widening of the distribution of $\bar{Q}$ causes a widening in the distribution of Ξ, then observations of $\Xi < \xi_\alpha$ will happen at greater than the nominal rate. Similarly, as illustrated in figure 8.6, in confidence interval estimation, the same widening can cause the true performance to be excluded from an interval at a rate greater than the nominal α. For both significance tests and confidence intervals, the result is an increase in the type I error rate, resulting in permissive inference.

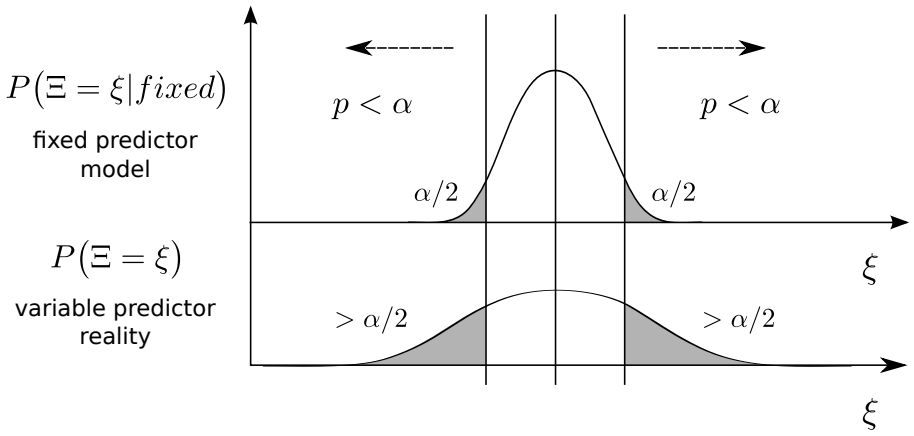

Figure 8.5: An illustration of how a wider than modelled distribution of a test statistic can result in an inflated result of false positives. The distribution of Ξ under a fixed predictor model is shown above. This is used to map observations of Ξ to p-values. The shaded areas of the distributions represent the areas where an observation of Ξ produces a $p \leq \alpha$. When this model is applied in a CV experiment with variable predictors, values of Ξ inducing p values of less than α under the fixed predictor model occur at a greater than expected rate ($> \alpha$) even when the marginal mean is the same. Thus, when a fixed predictor model is applied to test against a null hypothesis $\mu_u = \eta$, the rate of false positives will be higher than the nominal α.

In certain cases, the relationships between the $Q_{r,j}$ may actually lead to a narrower distribution for $\bar{Q}$ than expected under a fixed predictor assumption, making inference conservative. This is not normally regarded as a problem, as it does not undermine the validity of statistical analysis.

8.3.1 Asymptotic correctness of the fixed predictor models

As the size l of a random training set $\mathbf{G}$ grows, the random variability of the predictor $T = u(\mathbf{G})$ constructed by some learner u will decrease. As l becomes sufficiently large, T will produce predictions that are arbitrarily close to those of some constant t with overwhelming probability. In this limit, the random variability of T can be neglected, and the fixed predictor model is asymptotically correct. Exactly how large l has to be for this to occur will depend on the

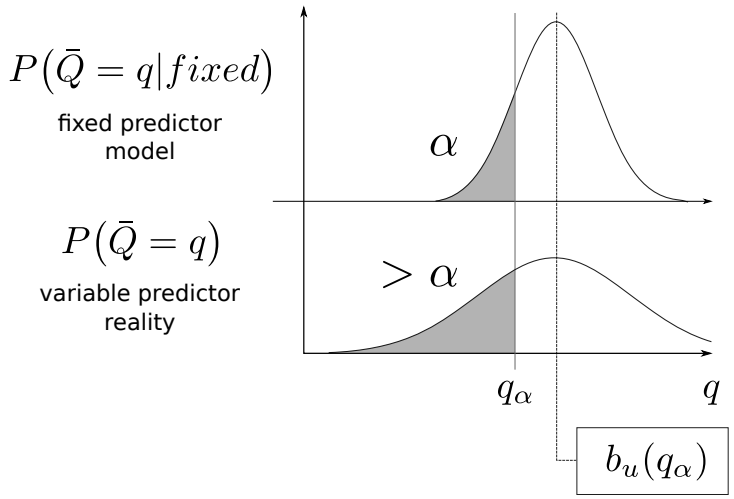

Figure 8.6: This diagram considers a one-sided interval procedure in which the boundary b_u is a function of the observed performance $\bar{Q}$ alone. Any observation $\bar{Q} < q_\alpha$ will lead to a selected bound $b_u(\bar{Q})$ that excludes the learner performance μ_u. By design, this only occurs with a probability of α under the fixed predictor assumption (above). However, when the same assumption is applied in a CV experiment with variable predictors, the wider distribution of $\bar{Q}$ causes the μ_u to be excluded from the interval at a rate greater than α. The interval will then contain μ_u at less than the nominal rate of $(1-\alpha)$.

precise characteristics of the learning problem, so it is hard to produce a general rule to suggest when the fixed predictor model can be applied. It is intuitively unlikely that this limit should be reached in ANA research; high dimensional feature spaces mean that many parameters are required to specify a predictor, and this generally makes predictor selection less stable.

8.3.2 Unknowable distribution form

It is not only the level of correlation between the component results that is unknown; the exact form of their marginal joint distribution is unknown too. This in turn means that the form of the distribution of $\bar{Q}$ is unknown, and this poses a problem for statistical analysis.

There are specific cases where there are reasons that $\bar{Q}$ should be normal. In the limit where l is large and predictor selection is stable, the performance results on the distinct items of the dataset **D** will be independent, and an average of them will be normally distributed. For LPOCV and RKCV or RHOCV with very large numbers of repetitions, in the limit where $l \to \infty$ while the training set size m stays fixed, $\bar{Q}$ must be asymptotically normally distributed [135]. It seems unlikely that this limit will be approached in ANA research.

In the general case, one must pragmatically proceed with the normal and binomial models that one knows cannot be strictly justified. This can still produce procedures with reasonable empirical performance. Crucially, in most cases, there is no alternative.

8.4 Performance measures in a hold-out experiment

The fixed predictor models describe the case where pre-specified predictors are evaluated on a random test set $\mathbf{H}$ of i.i.d. items. In this context, a predictor t and its performance μ_t may be taken as constants and the Q_i are i.i.d. random variables with marginal mean μ_t. In a hold-out experiment with a learner u, there is also a random training set $\mathbf{G}$ of m i.i.d. items from which a predictor $T = u(\mathbf{G})$ is produced. The predictor is now a random variable with an unknown distribution, as is its performance M_T. The Q_i are defined

$$Q_i = \phi\big(T(X_i), Y_i\big) \text{ where } H_i = (X_i, Y_i)\,, \tag{8.27}$$

and the fixed predictor model now describes the distribution of the Q_i *conditional* on the even $T = t$; the Q_i are independent conditional on T with a shared conditional mean M_T . This means that the Q_i remain identically distributed, but they are *no longer marginally independent*. The marginal mean of the Q_i is $\mu_u = \mathbb{E}[M_T]$ and the distribution of an individual Q_i is now the mixture

$$P(Q_i = q) = \int_t dt\, P(Q_i = q \,|\, T = t) P(T = t)\,, \tag{8.28}$$

where $P(Q_i = q \,|\, T = t)$ is the fixed predictor distribution for the Q_i produced by a predictor t.

The change in the distribution of the Q_i will affect the distribution of derived test statistics. In general, one should expect the marginal variance of the Q_i to be greater when the predictor performance is variable rather than fixed; by the law of total variance

$$\begin{aligned} \mathbb{V}\text{ar}[Q_i] &= \mathbb{E}\big[\mathbb{V}\text{ar}[Q_i \,|\, T]\big] + \mathbb{V}\text{ar}[\mathbb{E}[Q_i \,|\, T]] \\ &= \mathbb{E}\big[\mathbb{V}\text{ar}[Q_i \,|\, T]\big] + \mathbb{V}\text{ar}[M_T]\,. \end{aligned} \tag{8.29}$$

The term $\mathbb{E}\big[\mathbb{V}\text{ar}[Q_i \,|\, T]\big]$ may be taken as loosely representing the variance expected under a fixed predictor model where $M_T = \mu_u$, while the term $\mathbb{V}\text{ar}[M_T]$ may be read as a strictly positive addition due to the variability of T. This holds true for both the case where a single predictor is being studied and the case where two are being compared.

More can be said about the precise distribution of test statistics under the specific models.

8.4.1 Under the normal model

Under the normal model, in the case of a fixed predictor, each of the Q_i could be modelled

$$Q_i = \mu_t + \sigma_t \varepsilon_i\,, \tag{8.30}$$

where the ε_i are standard normal variables and σ_t is a positive constant. In a single hold-out experiment with random predictor T, the predictor performance M_T may be modelled as being equal to the learner performance μ_u plus some normal zero-mean perturbation term d_T. The Q_i must be modelled

$$\begin{aligned} Q_i &= M_T + \sigma_T \varepsilon_i \\ &= \mu_u + d_T + \sigma_T \varepsilon_i \,, \end{aligned} \tag{8.31}$$

where σ_T is a random variable describing standard deviation of the Q_i conditional on T. Both of these have unknown distributions. Under the simplifying assumption that σ_T does not change with T, and that d_T has a normal distribution with variance $\sigma_{d_T}^2$, the form of the derived test statistics' distributions does not change; $\bar{Q}$ is normally distributed with mean μ_u and variance $\sigma_T^2 + \sigma_{d_T}^2$. The distribution of S^2 is unaffected by the introduction of a variable d_T, and S^2/σ_T^2 remains χ^2 distributed with $n-1$ DOF. This makes S^2/n a downwardly biased estimator of the variance of $\bar{Q}$, and it means that the distribution of the statistic $n\bar{Q}/S^2$ is a *rescaled* t-distribution, which is wider by a factor of

$$\sqrt{\frac{n\sigma_{d_T}^2 + \sigma_T^2}{\sigma_T^2}} \,. \tag{8.32}$$

Because both Z and t-statistics have a distribution that is wider (by the scaling factor) than expected under the fixed predictor model, interval and significance testing procedures based on fixed predictor assumptions will have elevated type I error rates.

8.4.2 Under the binomial model

In the binomial model, as under the normal model, the distribution of $\bar{Q}$ becomes wider. By the law of total variance

$$\begin{aligned} \mathbb{V}\mathrm{ar}[n\bar{Q}] &= \mathbb{E}\big[n\mathbb{V}\mathrm{ar}[\bar{Q}\,|\,T]\big] + \mathbb{V}\mathrm{ar}\big[\mathbb{E}[n\bar{Q}\,|\,T]\big] \\ &= \mathbb{E}\big[\mathbb{V}\mathrm{ar}[n\bar{Q}\,|\,T]\big] + \mathbb{V}\mathrm{ar}[nM_T] \\ &= \mathbb{E}\big[nM_T(1-M_T)\big] + n^2\mathbb{V}\mathrm{ar}[M_T] \\ &= \Big(n\mu_u - n\mathbb{E}[M_T^2]\Big) + \Big(n^2\mathbb{E}[M_T^2] - n^2\mu_u^2\Big) \\ &= n\mu_u - n\mu_u^2 + n(n-1)\mathbb{E}[M_T^2] - n(n-1)\mu_u^2) \\ &= n\mu_u - n\mu_u^2 + n(n-1)\mathbb{V}\mathrm{ar}[M_T] \end{aligned} \tag{8.33}$$

From this, one may subtract the variance

$$\mathbb{V}\mathrm{ar}[n\bar{Q}\,|\,M_T = \mu_u] = n\mu_u - n\mu_u^2 \,. \tag{8.34}$$

expected under the fixed predictor model in which $M_T = \mu_u$, to produce the difference

$$\begin{aligned} \mathbb{V}\text{ar}[\bar{Q}] - \mathbb{V}\text{ar}\left[n\bar{Q} \,|\, M_T = \mu_u\right] &= n(n-1)\mathbb{V}\text{ar}[M_T] \\ &\geq 0\,, \end{aligned} \tag{8.35}$$

thus demonstrating that the change in variance due to M_T's variability is strictly non-negative.

As under the normal model, the fact that $\bar{Q}$'s distribution is wider than expected under a fixed predictor assumption will cause inference based on that assumption to become permissive.

8.4.3 Interval coverage reduction

The following is an intuitive explanation for why interval procedures based on fixed predictor models will have reduced coverage in SHOCV. Consider the rate at which a general value η is contained in a confidence interval conditional on the value of the predictor performance M_T. In general, this will be $(1-\alpha)$ when $\eta - M_T = 0$, and will decrease monotonically with $|\eta - M_T|$. This conditional containment rate is illustrated for the case of the binomial model in figure 8.7. Where $|\mu_u - M_T|$ is non-zero with some non-negligible probability, then μ_u must be contained in the interval with a rate of less than $(1-\alpha)$.

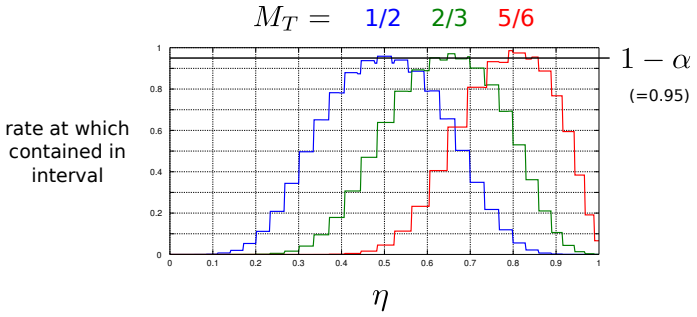

Figure 8.7: Rates at which points are contained in confidence intervals in SHOCV with $n = 25$ under the binomial model. Various possible instances of M_T are illustrated in different colours. It can be seen that the containment rate drops from $(1-\alpha)$ when $|M_T - \eta|$ increases from 0.

8.5 Performance measures in a general cross validation experiment

In the single hold-out experiment, the Q_i have a relatively simple relationship to one another; they are conditionally independent given the performance M_T of the single random predictor constructed. In a more general CV experiment with multiple train-test experiments, there are a variety of factors controlling the joint distribution of the item performance measures. While the item performance measures within a single test set will have the same joint distribution as those in the test set of a single hold-out experiment, there will be relationships between item performance measures in different test sets that are found only in more complex CV experiments.

This section will discuss the various factors contributing to dependency between the $Q_{r,j}$ in a general CV experiment, and describe how the index sets used to define two train-test experiment will affect the joint distribution of their results. In addition to this general description, two specific models are provided: one for KCV and one for CV experiments involving repeated measurements derived from random partitions.

8.5.1 Factors controlling the joint distribution of performance results

Recall that a CV experiment on a set of items $\mathbf{D} = \langle D_i \rangle_{i=1}^{l}$ is defined by a design $\mathbf{I} = \langle I_r \rangle_{r=1}^{R}$. I_r is a set of m indices in the range $\{1,2,...,l\}$ that defined the rth training set $\mathbf{G}_r$, and $J_r = \{1,2,...,l\} \setminus I_r$ is a set of $n = l - m$ indices defining the rth testing set. Let

$$J_r = \{\imath_{r,1}, \imath_{r,2}, \cdots, \imath_{r,n}\}$$

denote the indices of the rth testing set, $T_r = u(\mathbf{G}_r)$ denote the predictor selected on the rth training sets, and $D_i = (X_i, Y_i)$ denote the ith item of the full sample $\mathbf{D}$. The performance measure $Q_{r,j}$ on the jth item of the rth testing set may now be defined

$$Q_{r,j} = \phi\left(T_r(X_{\imath_{r,j}}), Y_{\imath_{r,j}}\right).$$

Consider a CV strategy in which all training and testing sets have the same marginal distribution as those of a commensurate hold-out experiment. For all r and j, $Q_{r,j}$ will have the same marginal distribution as some Q_i in the hold-experiment. For a given r, the $Q_{r,j}$ will have the same joint distribution as the Q_i. However, for any j, j', r and $r' \neq r$, $Q_{r,j}$ and $Q_{r',j'}$ may have a joint distribution not found between the Q_i of the hold-experiment.

As discussed in [135], the joint distribution of $Q_{r,j}$ and $Q_{r',j'}$ will depend on the following:

1. whether $\imath_{r,j} = \imath_{r',j'}$,
2. $|I_r \cap I_{r'}|$, and
3. whether $\imath_{r,j} \in I_{r'}$ and/or $\imath_{r',j'} \in I_r$.

The reasons these three factors affect the joint distribution of the performance measures can be intuitively explained as follows:

1. Reuse of items for testing. While the predictors produced on the various training sets in a CV experiment will vary, if the learner is truly able to identify any relationship between the features and the labels, they can be expected to share some common aspects. This means that the labels of items can be persistently easy or difficult to predict. When

a given item D_i from the full dataset is used for performance evaluation in two different train-test experiments, the results are likely to be similar. Thus, the performance measures of different train-test experiments on D_i will be correlated. When the predictor may be taken as fixed over the full CV experiment, the results for D_i are constant over all train-test experiments, and the correlation is perfect.

2. **Shared training items.** Items can have a persistent effect on the quality of the predictor they produce when they appear in a training set. When the training sets of two train-test experiments share items, the performances of the predictors produced on them are more likely to be similar than those if they shared no items. Thus, the predictor performances are dependent. Through their dependency on the respective predictor performances, the performance measures on the associated testing sets will be correlated with one another. The more the training sets have in common, the greater the correlations are likely to be.

3. **Train-test overlap.** In KCV with $K = 2$, there are no items that are present in either both of the training sets or both of the testing sets, and the results in each testing set might appear to be mutually independent. However, this is not the case [40]. As already mentioned, items can have a persistent effect on the performance measure produced in a train-test experiment when they appear in either the training or testing set. The persistent training and testing effects that items have may be related. For instance, an item that has a very atypical relationship between its features and labels will be hard to make a correct prediction for, and may also lead to a poorer selection of predictor. In KCV with $K = 2$, the presence of these items in the training set of the first train-test experiment heralds a low performance estimate in both the first train-test experiment and the second. Conversely, in some classification problems, some 'borderline' items may be difficult to make predictions for, but may be very informative in the construction of predictors. This will mean that bad results in one test set herald good results in the other. In either case, there is a common contributory factor to both results, and they are no longer independent.

The factors determining the joint distributions of the item performance measures $Q_{r,j}$ in turn define the factors that determine the joint distribution of the average measures observed on each test set. Where

$$\bar{Q}_r = |J_r|^{-1} \sum_{j=1}^{j=|J_r|} Q_{r,j}, \tag{8.36}$$

the joint distribution of $\bar{Q}_r$ and $\bar{Q}_{r'}$ will be determined by $|I_r \cap I_{r'}|$, $|I_r \cap J_{r'}|$, $|I_{r'} \cap J_r|$, and $|J_r \cap J_{r'}|$. Where $J_r = \{1, 2, ..., l\} \setminus I_r$ and $|J_r| = n$ for all r, any one of these things is sufficient to determine

the others. An illustration of these possible set overlaps is provided in figure 8.8.

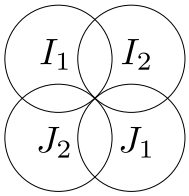

Figure 8.8: Possible set membership of items in two train-test experiments.

A precise description of the covariance between the performance results in a minimal regression problem with normally distributed features and labels is presented in the supplementary material of [135]. In that document, $\tau_2^{(3)}$ represents the covariance between item performance estimates in the different test sets of KCV with $K = 2$. That it is non-zero demonstrates that *the results on the separate test sets are not independent even in this case*.

8.5.2 K-fold cross validation

Let $Q_{k,j}$ denote the performance measured on the jth item out of n in the kth testing set out of K. A consideration of the symmetries between the performance results on the different items of a KCV experiment yields the pairwise dependency structure described in [38]; the relationship between two different item performance measures depends solely on whether their associated items are in the same testing set. Under a normal model, the joint distribution of the $Q_{r,j}$ is determined solely by their covariance, as they share the same marginal mean. The three pairwise covariances, denoted A, B and C, define the full covariance matrix

$$\mathbb{C}\text{ov}[Q_{k,j}, Q_{k',j'}] = A\delta_{kk'}\delta_{jj'} + B(\delta_{jj'} - \delta_{kk'}) + C(1 - \delta_{kk'}), \tag{8.37}$$

where $\delta_{i,j}$ represents the Kronecker delta. This is illustrated in figure 8.9.

Let $\bar{Q}_k$ denote the mean observed performance on the kth test set, and $\bar{Q}$ denote the mean observed performance over all items. It is straightforward to show that

$$\mathbb{V}\text{ar}[\bar{Q}_k] = \frac{A + (n-1)B}{n}, \tag{8.38}$$

$$\mathbb{C}\text{ov}[\bar{Q}_k, \bar{Q}_{k' \neq k}] = C, \text{ and} \tag{8.39}$$

$$\mathbb{V}\text{ar}[\bar{Q}] = \frac{A + (n-1)B + (l-n)C}{l}. \tag{8.40}$$

The three constants A, B and C are unknown, and will vary depending on K, l and the learning problem under study. Of these, A is positive by its definition as a variance, while B represents the covariance between the items of a train-test experiment with m items in the

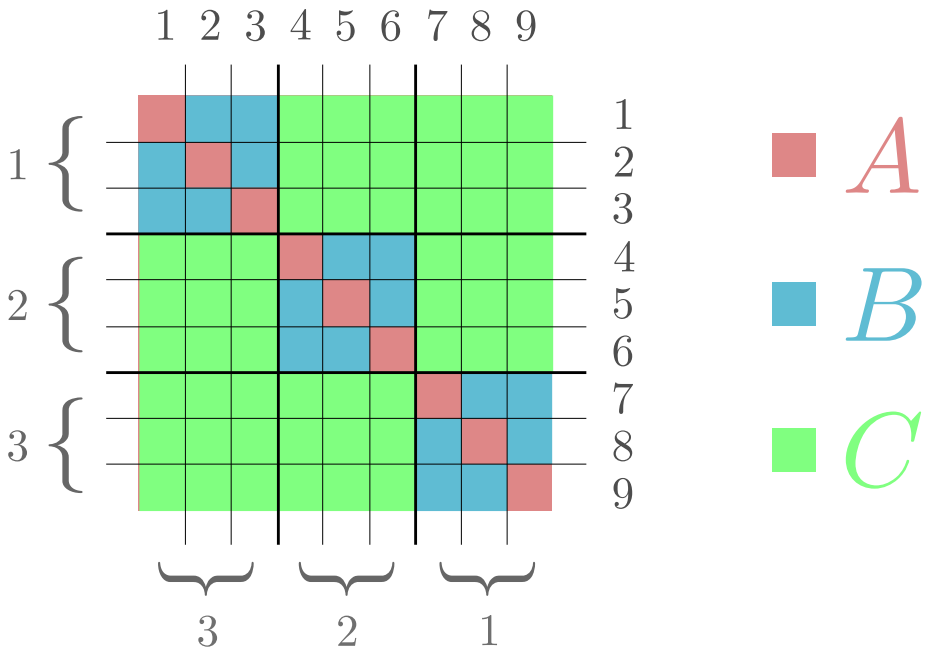

Figure 8.9: Illustration of the three covariance values in KCV where $l = 9$, $K = 3$ and $n = 3$. Brackets denote testing set membership, with bracket k indicating item indices in ${}_H I^{(k)}$.

training set, which means that $B \leq A$. If B was negative, a sufficiently high value of n in a single hold-out experiment would cause the mean performance to have a negative variance according to equation (8.38); thus $B \geq 0$. As for C, it has most commonly been measured as positive in most practical tasks, though negative values are possible [178]. C will typically increase with K [35, 36], as the training sets of KCV will share more items, making the resulting predictors more similar.

Normal model inference based on a fixed predictor assumptions may assume either that the $\bar{Q}_k$ are independent, or that all the $Q_{k,j}$ are. In either case, the precise bias of the resulting variance estimate, and thus the additional width of any Z or t-statistic will depend on the values of the constants A, B and C. If C takes a negative value of sufficiently large magnitude, it is even possible that fixed predictor inference will be conservative. Most commonly however, one should expect that variance estimates will be somewhat negatively biased, making fixed predictor inference moderately permissive.

As will be discussed in chapter 9, though it might be more intuitively plausible than in SHOCV, it is not possible to estimate the variance of $\bar{Q}$ in KCV without bias.

8.5.3 Repeated experiments using random partitions

In both RKCV and RHOCV, E random partitions are used to produce multiple performance estimates. Let $\bar{Q}_e$ denote the performance estimate from the eth random partition, and $\bar{Q}$ denote the mean over all repetitions. If the $\bar{Q}$ are assumed to be jointly normal, then the only statistics required for inference are mean and variance estimations.

Where the covariance between the $\bar{Q}_e$ is given

$$\mathbb{C}\text{ov}[\bar{Q}_e, \bar{Q}_{e'}] = \sigma^2 \delta_{ee'} + \rho(1 - \delta_{ee'}), \tag{8.41}$$

where $\rho < \sigma^2$, the variance of the grand mean over all partitions is

$$\mathbb{V}\text{ar}[\bar{Q}] = \frac{\sigma^2 + (E-1)\rho}{E}. \tag{8.42}$$

Because this must be strictly positive for all n, $\rho \geq 0$.

It is straightforward to show that the naive variance estimator

$$S^2 = \frac{1}{E(E-1)} \sum_{e=1}^{E} (\bar{Q}_e - \bar{Q})^2 \tag{8.43}$$

for $\bar{Q}$ has an expectation of

$$\mathbb{E}[S^2] = \frac{\sigma^2 - \rho}{E} \tag{8.44}$$

and bias of precisely $-\rho$. As $E \to \infty$, the variance estimator S^2 will always converge to zero, while the true variance of $\bar{Q}$ will converge to ρ.

When the t-statistic $E\bar{Q}/S$ is used, the negative bias of the variance estimator will cause the resulting inference to become permissive. Because the ratio of the true variance to the estimated variance will increase with E, larger values of E will make the test ever more permissive. This behaviour has been observed in the 'resampled t-test', a normal model test assuming independence between sequential partition results in RHOCV [40, 179]. As E becomes arbitrarily large, arbitrarily small mean observed differences in performance can become significant; the probability of false positive in a one-sided test will approach $1/2$, while the probability in a two-sided test will approach 1. *Any inference assuming independence between the results on sequential random partitions of a give dataset is therefore a very bad idea.*

8.5.3.1 Insight from the fixed predictor model

Notably, the assumption of independence between the sequential $\bar{Q}_e$ implicit in the naive variance estimator cannot even be justified by the fixed predictor model, where sequential obser-

vations $\phi(t(X_i), Y_i)$ on an item $D_i = (X_i, Y_i)$ from a full dataset $\mathbf{D} = \langle D_i \rangle_{1 \leq i \leq l}$ will always produce the same result. This means that, in RKCV, the $\bar{Q}_e$ should all be perfectly correlated. In RHOCV, the $\bar{Q}_e$ would have a correlation equal to n/l, the expected fractional overlap of the testing sets [37].

8.6 Replicability and repeatability

In this section, I shall formally introduce the ideas of repeatability and replicability in an inference context and explain their relationship with the error rates.

Replicability is defined as the chance that two experimenters enacting the same experiment with *the same* data will reach the same conclusion. Low replicability implies a high level of 'internal randomness' in a particular experiment. High replicability allows experiment repetitions to identify fraud and experimental error. Repeatability is the chance that two experimenters enacting the same experiment with *different* data will reach the same conclusion. High repeatability implies that *all* random effects have only a small role in determining the outcome of an experiment. This includes both the 'internal' components and those relating to random production of the dataset.

Consider a statistical procedure producing a binary result $Y \in \{0, 1\}$. Y could be the non-containment of a parameter value in a confidence interval or the rejection of a null hypothesis. This procedure takes a random dataset $\mathbf{D}$ as input, but has some internal randomness related to the random generation of a CV design. After a given dataset has been observed, a potentially infinite sequence of experiments may be conducted, producing the sequence of binary results $\langle Y_e \rangle_{e=1}^{E}$. The Y_e are exchangeable random variables, so they may be regarded as being i.i.d. conditional on the dataset [180]. Because the Y_e are binary, they must have a Bernoulli distribution with a random variable rate parameter W specified by the dataset. In the case of a CV experiment in which there are only a finite number of possible experiment designs that are all selected with equal probability, this W is the fraction of those designs which lead to an outcome $Y = 1$ when applied to the dataset D. The marginal probability that $Y = 1$ is $\mathbb{E}[W]$. This will be equivalent to the power of the test under the alternative hypothesis or the type I error rate under the null hypothesis.

The chance that two sequential results on the observed dataset will be equal conditional on $W = w$ is given $w^2 + (1 - w)^2$. Thus, the replicability, which is the probability that two

sequential results will be the same on a random dataset, is given

$$\begin{aligned} \text{replicability} &= \mathbb{E}[W^2 + (1-W)^2] \\ &= 2\,\mathbb{E}[W^2] - 2\,\mathbb{E}[W] + 1\,, \end{aligned} \tag{8.45}$$

The repeatability, which is the chance that two results on two independent datasets will be the same, is given

$$\begin{aligned} \text{repeatability} &= \mathbb{E}[W]^2 + (1-\mathbb{E}[W])^2 \\ &= 2\mathbb{E}[W]^2 - 2\mathbb{E}[W] + 1. \\ &= \text{replicability} - 2\big(\mathbb{E}[W^2] - \mathbb{E}[W]^2\big) \\ &= \text{replicability} - 2\mathbb{V}\text{ar}[W]. \end{aligned} \tag{8.46}$$

An effective statistical procedure will have an $\mathbb{E}[W]$ near 1 under an alternative hypothesis, and a $\mathbb{E}[W]$ near 0 under the null hypothesis. As can be seen in the first line of equation (8.46), this will require a high repeatability. As can be seen in the last line, repeatability is strictly limited by replicability. This demonstrates that high replicability is a necessary (though not sufficient) requirement on an effective statistical procedure.

Summary

Statistical inference for learner performance quantities on the basis of CV results must often be conducted using fixed predictor models that fail to account for the dependencies between component results. As a consequence of this, inference procedures for learner performances may often be permissive; significance tests may have type I error rates above the nominal α, and confidence intervals may have coverages less than the nominal $(1-\alpha)$. The degree to which this happens is unknown, and will depend on the precise specification of the CV strategy and learning problem under study. Inference based on an assumption of independence between the sequential experiments of EKCV and RHOCV is inadmissible. High repeatability and replicability are necessary requirements for an effective statistical procedure.

Chapter 9

Specialist inference for cross validation

This chapter is concerned with specialist statistical procedures for cross validation (CV). In it, I shall review various procedures from the literature and discuss their various strengths and weaknesses. I shall then review the practice of performance inference in the AD ANA research field and make recommendations for how it can be improved.

Much of the discussion of methods based on t-like statistics will focus on statistical testing. However, it is to be understood that these methods can be extended straightforwardly to produce confidence intervals.

9.1 Dietterich's approximate statistical tests

Dietterich's paper of 1998 was one of the first to discuss the problem of dependency [40]. In it, he considers the following five CV-based significance tests for a difference in performance between two learners:

- a resampled t-test, in which the test set measures $\bar{Q}_e$ from the sequential train-test experiments of RHOCV are treated as i.i.d. normal,
- a cross validated t-test, in which the $\bar{Q}_k$ of KCV with $K = 10$ are treated as i.i.d. normal,
- a McNemar's test, in which the item measures Q_i in SHOCV are treated as i.i.d. binomial,
- a two proportions test, in which the Q_i in SHOCV are treated as i.i.d. normal, and
- the 5×2 t-test, a new invention of Dietterich.

The 5×2 t-test is based on RKCV with $E = 5$ and $K = 2$. Where $\bar{Q}_{e,k}$ the mean performance on the kth test set in the eth KCV repetition of RKCV, the relevant test statistic is τ, defined

$$\tau = \frac{\bar{Q}_{1,1} - \mu}{\bar{S}^2}, \text{ in which}$$
$$\bar{S}^2 = \frac{1}{5}\sum_{e=1}^{5} S_e^2, \text{ where } S_e^2 = \frac{1}{2}(\bar{Q}_{e,1} - \bar{Q}_{e,2})^2.$$

The $\bar{Q}_{e,k}$ are assumed to be i.i.d. normal with some variance σ^2. This implies that the denominator $\bar{S}$ is χ distributed with the same scale parameter. This should give the statistic a t distribution with 5 DOF. The choice of $K = 2$ is made to reduce the correlation between the test sets results of KCV, and the choice of 5 is an empirically chosen compromise between the power of the test and the accuracy of the distribution model.

Dietterich evaluates the tests in a series of simulated classification problems. The results demonstrate that both the cross validated t-test and the resampled t-test can have unacceptable type I error rates. The resampled t-test is particularly unreliable, as its type I error rises dramatically with the number of train-test repetitions. The two tests based on SHOCV have acceptable type I error rates but relatively low power. The 5×2 t-test is advanced as a good compromise.

Remarks

The 5×2 t-test has greater power than the SHOCV tests because multiple component train-test experiments provide more stability in variance estimation. Importantly, it has lower error rates than the resampled t-test because its variance estimator is not strongly negatively biased by correlations caused by the reuse of items for testing.

The 5×2 t-test has been observed to be permissive in some contexts and to have lower power than similar alternatives [37, 179]. The permissive behaviour has two potential origins.

- The unmodelled covariance between $\bar{Q}_{e,1}$ and $\bar{Q}_{e,2}$ due to train-test overlap (see section 8.5.2) will bias the variance estimator.
- Due to the reuse of items in KCV repetitions, the variance estimates S_e^2 are not independent. The effect of this can be viewed as an effective reduction in the DOF in the distribution of $\bar{S}^2$ [179].

Both of these effects will make the distribution of τ wider.

If more KCV repetitions were used, this would provide a more stable variance estimator, increasing power. However, it would also increase the underestimation of the DOF of $\bar{S}$, causing the test to become permissive. The choice of 5 in [40] was a compromise between these two concerns based on preliminary calibration experiments. This choice will not be optimal for all problems, and this may be the reason for the test's poorer behaviour elsewhere [37, 179].

As pointed out in [37], even under the key assumption that the $\bar{Q}_{e,k}$ are i.i.d., the numerator and denominator of the test statistic are not independent. It is not clear whether this will make the test more or less permissive.

Another undesirable feature of the 5×2 t-test is the use of a single $\bar{Q}_{e,k}$ in the numerator. Alpaydin refined the 5×2 t-test to produce the 5×2 F-test [181] which makes use of all the

$\bar{Q}_{e,k}$ in the numerator and appears to have better behaviour. However, it does not avoid the two issues causing permissive behaviour, and it retains the issue of dependency between the numerator and denominator of its test statistic even under simplifying assumptions.

9.2 Modelling the covariance in RHOCV

In [37], Nadeau and Bengio consider the problem of inference for the learner performance using RHOCV. They develop a model (related to those of chapter 8) to describe the covariance between the item performance results $Q_{r,j}$ of RHOCV. They show that, without one of several simplifying assumptions, *it is not possible to estimate the variance of grand mean performance estimate $\bar{Q}$ without bias as a function of the $Q_{r,j}$ alone*. Consequentially, inference must proceed with a biased estimator based on some incorrect simplifying assumption.

They consider the resampled t-test, which effectively assumes all the $\bar{Q}_e$ are independent. Its test statistic,

$$\tau = \frac{\bar{Q}-\mu}{S^2} \text{ where } S^2 = \frac{1}{E}\cdot\frac{1}{E-1}\sum_{e=1}^{E}(\bar{Q}_e-\bar{Q})^2\,, \tag{9.1}$$

is assumed to have a t-distribution with $E-1$ DOF. As discussed in section 8.5.3, the variance estimator S^2 has a negative bias that becomes more severe with increasing E. If the $\bar{Q}_e$ are instead modelled as being jointly normal with a correlation ρ, the following modification will produce an unbiased estimator:

$$\tau' = \frac{\bar{Q}-\mu}{S'^2} \text{ where } S'^2 = \left(\frac{1}{E}+\frac{\rho}{1-\rho}\right)\cdot\frac{1}{E-1}\sum_{e=1}^{E}(\bar{Q}_e-\bar{Q})^2\,. \tag{9.2}$$

Nadeau and Bengio propose the **corrected resampled t-test**, in which ρ in equation (9.2) is the n/l. This is derived from the assumption that the correlation between two test set results is simply the fraction of overlap between them, as would be expected under a fixed predictor model.

Recognising that this assumption neglects other sources of correlation (see section 8.5.1), they propose an additional procedure in which the variance is estimated empirically. This is the **conservative Z-test**, and it can be used with arbitrary CV strategies. It is conducted as follows. First, a CV experiment is conducted on the full dataset of l items to produce a final performance measurement $\bar{Q}$. Then full dataset is randomly divided into two disjoint subsets of size $l/2$, and the same CV experiment (with training set sizes reduced by a factor of 2) is conducted on each of the disjoint subsets to produce two independent measurements $\bar{Q}'_1$ and $\bar{Q}'_2$. These are combined to produce $(\bar{Q}'_1-\bar{Q}'_2)^2/2$. By construction, this is an unbiased estimator of the variance of a CV performance estimate derived from an experiment on $l/2$ items. The random division and

variance estimation procedure is repeated c times, and the c resulting variance estimates are averaged together to produce the stabilised estimator $\hat{S}^2_{l/2}$. The test statistic

$$Z = \frac{\bar{Q} - \mu}{\hat{S}_{l/2}}. \tag{9.3}$$

is modelled to have a standard normal distribution. Because more items should lead to more stable performance estimation, $\hat{S}^2_{l/2}$ should be an upwardly biased estimator of the variance of $\bar{Q}$, and inference based on this statistic should be conservative. Preliminary experiments suggest that c should take a value of around 10.

After introducing their two new procedures, Nadeau and Bengio conduct a series of learner comparison experiments to examine the powers and type I error rates of the proposed tests and several alternatives. These include the resampled t-test and the 5×2 t-test with RHOCV, McNemar's test with SHOCV, and various related bootstrap tests. The experiments demonstrate the efficacy of the correction factor for the resampled t-test, which reduced the type I error from very high values (e.g., 0.6) to close to the nominal value (e.g., 0.12 for $\alpha = 0.1$). The corrected tests have better power than the 5×2 t-test with lower type I error rates. As intended, the conservative Z-test is conservative in all experiments.

Remarks

The variance estimate S'^2 appearing in the corrected resampled t-test has the desirable property that its bias is limited. While there is some bias due to the neglect of factors 2. and 3. in section 8.5.1, the ratio $\mathbb{E}[S'^2]/\mathbb{V}\mathrm{ar}[\bar{Q}]$ does not become arbitrarily large with increasing E. This is appealing, because it could allow the test to be used with low variance, high repetition RHOCV.

There is, however, a reason why this test may be permissive even under a fixed predictor model. Let t denote a fixed predictor. Where the performance measures $Q_i = \phi(t(X_i), Y_i)$ on the items $D_i = (X_i, Y_i)$ in the full dataset $\mathbf{D} = \langle D_i \rangle_{1 \leq i \leq l}$ are i.i.d. normal, the $\bar{Q}_e$ will indeed be marginally normal with a pairwise correlation of n/l, but they will not be *jointly normal*[1]

As can be confirmed by numerical experiment or theoretical consideration, in the limit $E \to \infty$, the variance estimator S^2 will become proportional to the quantity

$$\sum_{i=1}^{l} Q_i - \bar{Q}.$$

This means that S^2 will have a scale χ^2 distribution with $l-1$ DOF, rather than the much larger

[1] To illustrate this, consider the following: there are a total of $\binom{l}{n}$ possible testing sets, corresponding to $\binom{l}{n}$ possible $\bar{Q}_e$ values. Once all of these have been observed, an additional train-test experiment must take one of the observed values. This would not be the case if the $\bar{Q}_e$ were jointly normal.

$E-1$ of the model. In the high E limit, the model will therefore overestimate the DOF in the t-distribution of τ', leading to permissive inference. Good behaviour for relatively low values of l may therefore be contingent on a low value of E. As a minor improvement, I therefore suggest that $l-1$ be used as the DOF when $E > l$.

That the conservative Z-test allows it to be used with more efficient strategies than RHOCV is a desirable feature. Its variance estimator, $\hat{S}^2_{l/2}$, is the only one considered so far that is able to account for all sources of dependency. Accordingly, the test is the only one considered so far that will not be negatively affected by strong dependencies not expected under the fixed predictor model.

There is a flaw in the test made apparent by choice of the parameter c. In the preliminary experiments that suggested the value 10, it was observed that lower values lead to permissive inference, and higher values lead to unacceptably low power [37]. This shows that unmodelled variability in the variance estimator $\hat{S}^2_{l/2}$ is responsible for a large fraction of significant results. That is, if the $\hat{S}^2_{l/2}$ truly were stable, the test would be underpowered. To prevent this from occurring, low values of c are used to inject noise into the test statistic. Not only does this suggest that the power of the test is limited, but it also suggests that a correct choice of c is critical. Because the right choice of c will vary from problem to problem, any fixed value risks producing a permissive or underpowered test.

9.3 Modelling the variance of K-fold cross validation

Extending the earlier work of Nadeau and Bengio [37], Bengio and Grandvalet describe a model for the covariance structure of the item performance measures in KCV [38]. This is the same model described in section 8.5.2. To reiterate, where $Q_{k,j}$ represents the performance measured on the jth item of kth testing set, the covariance between two performance measures is given

$$\mathbb{C}\mathrm{ov}[\,Q_{k,j}, Q_{k',j'}\,] = A\delta_{kk'}\delta_{jj'} + B(\delta_{jj'} - \delta_{kk'}) + C(1-\delta_{kk'})\,. \tag{9.4}$$

The mean performance on the kth test set is denoted

$$\bar{Q}_k = n^{-1}\sum_{j=1}^{n} Q_{k,j}\,, \tag{9.5}$$

while the grand mean performance is given $\bar{Q} = \frac{1}{K}\sum_{k=1}^{K}\bar{Q}_k$. Consider a sequence of random variables $\langle X_i\rangle_{1\le i\le l}$ with a shared expectation and an average $\bar{X} = \frac{1}{l}\sum_{i=1}^{l} X_i$. Where S^2 denotes

the variance estimator

$$S^2 = \frac{1}{l(l-1)} \sum_i (X_i - \bar{X})^2 ,$$

the following identities hold

$$\mathbb{E}[S^2] = \frac{1}{l^2} \Big(\sum_i \mathbb{V}\text{ar}[X_i] - \frac{1}{l-1} \sum_{\substack{i \\ j \neq i}} \mathbb{C}\text{ov}[X_i, X_j] \Big) \tag{9.6}$$

$$\mathbb{V}\text{ar}[\bar{X}] = \frac{1}{l^2} \Big(\sum_i \mathbb{V}\text{ar}[X_i] + \sum_{\substack{i \\ j \neq i}} \mathbb{C}\text{ov}[X_i, X_j] \Big) \tag{9.7}$$

$$\mathbb{E}[S^2] - \mathbb{V}\text{ar}[\bar{X}] = -\Big(\frac{1}{l-1} + \frac{1}{l^2} \Big) \sum_{\substack{i \\ j \neq i}} \mathbb{C}\text{ov}[X_i, X_j] . \tag{9.8}$$

Using identity (9.7), it is straightforward to show that

$$\mathbb{V}\text{ar}[\bar{Q}_k] = \frac{A + (n-1)B}{n} , \qquad \mathbb{C}\text{ov}[\bar{Q}_k, \bar{Q}_{k' \neq k}] = C ,$$
$$\mathbb{V}\text{ar}[\bar{Q}] = \frac{A + (n-1)B + (l-n)C}{l} .$$

Bengio and Grandvalet show that, in the general case, *it is not possible to estimate the variance of $\bar{Q}$ without bias as a function of the item performance measures alone* [38]. This negative finding echoes the result of Nadeau and Bengio for RHOCV [37]. As in that case, inference using normal models must proceed using biased variance estimates derived from incorrect simplifying assumptions. Upwardly biased estimators will tend to produce conservative inference, while a downwardly biased ones will tend to produce permissive inference.

A great deal of effort has gone into finding estimators that are practically useful [182]. I shall now list and consider these. Where I describe their bias, I have computed this using identities (9.6) though (9.7).

1. The following estimator based on a fixed predictor model:

$$S_1^2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} (\bar{Q}_k - \bar{Q})^2 . \tag{9.9}$$

   This has a bias of precisely

$$\mathbb{E}[S_1^2] - \mathbb{V}\text{ar}[\bar{Q}] = -\Big(\frac{1}{k-1} + \frac{1}{k^2} \Big) C \tag{9.10}$$

   and a scaled χ^2 distribution with the $K-1$ DOF expected under a normal model. This results in what Dietterich calls the cross validated t-test.

2. A second estimator based on a fixed predictor model, defined

$$S_2^2 = \frac{1}{l(l-1)} \sum_{k=1}^{K} \sum_{j=1}^{n} (Q_{k,j} - \bar{Q})^2 , \tag{9.11}$$

has a bias of

$$\mathbb{E}[S_2^2] - \mathbb{V}\text{ar}[\bar{Q}] = -\Big(\frac{1}{l-1} + \frac{1}{l^2}\Big)\big(n^2 k(k-1)C + n(n-1)KB\big) \tag{9.12}$$

This is unbiased if both B and C are zero or C takes a negative value that exactly cancels out the effect of B. In this case, it will provide more powerful inference than S_1^2 under the normal model, as its scaled χ^2 distribution will have a full $l-1$ DOF. When it is not the case that $B = C = 0$, it will not have a scaled χ^2 distribution, but a scaled Gamma distribution. Briefly, this is because, though there exists a transform such that S_2^2 can be expressed as the norm of a normally distributed vector with independent elements, those elements have different scale parameters [183].

3. This estimator may be generalised using the assumption that C is characterised by $\mathbb{C}\text{ov}[\bar{Q}_k, \bar{Q}_{k' \neq k}]/\mathbb{V}\text{ar}[\bar{Q}] = a$, for some known a. This produces the estimator

$$S_3^2 = \frac{1}{1-a} \cdot \frac{1}{K(K-1)} \sum_{k=1}^{K} (\bar{Q}_k - \bar{Q})^2 \tag{9.13}$$

suggested by Grandvalet and Bengio in [178]. Grandvalet and Bengio suggest a conservative choice of $a = 0.7$ to ensure that type I error rates are likely to stay below their nominal values.

4. In [182], the authors describe how a variance estimator may be constructed using the properties of the learning problem the method of moments.

Remarks

Estimator 1. and the associated cross validated t-test is the 'default' option. As C is most commonly positive [178], this estimator will often produce somewhat permissive tests. Estimator 2. should generally be expected to be more biased than estimator 1. This is due to both the previously absent effect of B and a greater effect of C; a cursory inspection shows that the negative coefficient on C is of greater magnitude in the bias of estimator 2. than in that of estimator 1. unless $K = l$ and equality holds (see equations (9.12) and (9.10) respectively). The increased bias and the deviation from a χ^2 distribution under deviations from the single predictor model will make inference with the second estimator more permissive.

While a good choice of the parameter a in estimator 3. will produce well behaved inference, it is difficult to know how this is to be selected. If the conservative choice of a is made as suggested in [178], this will often lead to underpowered tests with their own harmful consequences [12]. It is difficult to imagine a reliable set of rules that would allow researchers to select a good a in advance.

Estimator 4. does not provide a general solution, as few researchers will have the time or ability to produce the deep mathematical models required for every new learner performance inference task they wish to consider.

9.4 Bouckaert's work on replicability

In a series of works between 2003 and 2005, Bouckaert and co-authors explore ideas of calibration and replicability in statistical tests for the comparison of learning algorithms [179, 184–187]. Estimating and increasing replicability, defined as the probability of getting the same outcome from a CV-based statistical test in two repetitions on a fixed dataset, is a key goal of these works. This is achieved primarily through the combination of performance results from sequential experiment repetitions in RHOCV and RKCV.

Let $\bar{Q}_{e,k}$ denote the mean result in the kth test set in the eth KCV repetition of RKCV. Let $\bar{Q}_{e,(k)}$ denote the kth *highest* test set result in the eth repetition. In [179], various combination rules are considered for t-like statistics, including

- the corrected and uncorrected resampled t-test and
- various t-tests for RKCV where one the following sets of measures are assumed to be i.i.d.:
  - $\bar{Q}_{e,k}$,
  - $\bar{Q}_{e} = K^{-1}\sum_k \bar{Q}_{e,k}$,
  - $\bar{Q}_{,k} = E^{-1}\sum_e \bar{Q}_{e,k}$, and
  - $\bar{Q}_{,(k)} = E^{-1}\sum_e \bar{Q}_{e,(k)}$.

None of these approaches have a strict theoretical grounding, and all of them can be permissive. To compensate for this, the DOF used to model the distribution of the relevant t-statistic is calibrated to ensure the nominal type I error rate is observed in a synthetic classification problem. The calibrated tests have better performance than their uncalibrated counterparts, and tests using more repetitions of RHOCV and RKCV are shown to have better replicability. The tests using RKCV are shown to have particularly good power.

Later work of Bouckaert's considers sign tests [184] and replicability in larger datasets [186]. It shows RKCV can provide more powerful and replicable tests than RHOCV

with a similar number of train-test experiments [185, 187].

Remarks

Bouckaert's work demonstrates that the use of multiple experiment repetitions can be useful and that the increased efficiency of RKCV over RHOCV can offer some benefit in inference. Unfortunately, the reliance on calibration can produce statistical tests that do not fare well in problems dissimilar to the ones used for calibration. As discussed in section 8.6, even if it is not a primary goal, replicability is a desirable feature of a statistical test.

9.5 An asymptotically correct U-statistic test

If a CV experiment is conducted on disjoint subsets of the same size, as occurs in the conservative Z-test of [37], this produces independent observations of the final performance estimate that can be used to estimate its variance without bias. This could in turn be used to estimate the variance of the average performance over all disjoint subsets to produce a t-like statistic. A similar idea is exploited in the asymptotically correct Z-test of Fuchs et al. [167], who show that for many type of strategy when $(m+1) \leq l/2$, the existence of independent item performance measurements allows one to produce an unbiased estimator for the final performance estimate [135, 188]. (This is true in KCV-like procedures such as EKCV where m falls in the required range, but not in true KCV.)

One key insight is that the grand performance measure $\bar{Q}$ of LPOCV can be viewed as a U-statistic [167]. Briefly, a U-statistic computed on a sample $\langle D_i \rangle_{1 \leq i \leq l}$ of i.i.d. random variables in $\mathbb{R}^d$ is one which takes the form

$$\binom{l}{g}^{-1} \sum_{s \in S} \psi(D_{s_1}, D_{s_2}, ..., D_{s_g}), \tag{9.14}$$

where $\psi : \mathbb{R}^{d \times g} \to \mathbb{R}$ denotes a symmetric function called a kernel, and S denotes the set of subsets $s = \{s_1, s_2, ..., s_g\}$ of the indices $\{1, 2, ..., l\}$ such that $|s| = g$. In the limit where $l \to \infty$, a U statistic with a given ψ will be asymptotically normally distributed [189]. The $\bar{Q}$ of LPOCV with a training set size of m may be viewed as a U-statistic in which the kernel function is LOOCV performance estimation on $m+1$ items. The unbiased variance estimator, denoted $\hat{V}$, is also a U-statistic. In the limit where l increases while m stays fixed, $\hat{V}$ will converge to the true variance of $\bar{Q}$, and $\bar{Q}$ itself will be normally distributed. By Slutsky's theorem, the test statistic $\hat{V}^{-1/2}(\bar{Q} - \mu)$ will have an asymptotically normal distribution, allowing an asymptotically correct Z test.

In practice, it is not necessary to conduct the exhaustive LPOCV needed to produce $\hat{V}$ and

$\bar{Q}$. Where performance measures are bounded, one can specify some E for RHOCV needed to ensure that approximations for these statistics will converge to the LPOCV case to within a certain tolerance [167, section 5].

Remarks

The limit where $l \to \infty$ while m stays fixed rarely occurs in practice; researchers normally wish to make use of the larger training sets afforded by large samples, so it is more reasonable to expect m to increase proportionally with l. This makes the assumption that $\bar{Q}$ is normal harder to justify. That said, this assumption is shared by all of the other tests considered, so it cannot be considered a particular flaw of this test.

A more serious potential problem with the test is the variability of $\hat{V}$ at finite sample sizes is neglected. In empirical demonstrations in problems related to learner selection the variability of $\hat{V}$ was observed to be so large that negative values occurred frequently [188]. This unmodelled variability could lead to permissive behaviour. There is also no guarantee of the independence of $\bar{Q}$ and $\hat{V}$, though it is unclear what the effect of that would be. There has as yet been no convincing empirical demonstration of the error rates associated with this test.

9.6 Non-parametric methods

This section will consider two non-parametric methods: the permutation test and the bootstrap.

9.6.1 Permutation testing

Let $\mathbf{D} = \langle D_i \rangle_{i=1}^{l}$, where $D_i = (X_i, Y_i)$, denote the full sample of items. Let $\mathbf{A} = \langle a_i \rangle_{i=1}^{l}$ denote a permutation of the integers $\{1, 2, ..., l\}$. Let $\mathbf{D}'$ in which $D'_i = (X_{a_i}, Y_i)$ denote the permuted sample induced by $\mathbf{A}$. Under the null hypothesis that there is no relationship between the features and the labels, there is no reason that any permutation should produce a dataset that leads to a higher performance measurement in a CV experiment than any other. Under this null hypothesis, if one conducts a CV experiment first on the original sample $\mathbf{D}$ and then on $N-1$ modified samples produced by random permutations, the probability that the measurement derived from the original ordering is in the top M measurements derived is precisely M/N. Where the original sample measurement ranks Mth out of N, this allows M/N to be treated as a p value. When one wishes to test against the null hypothesis that a learner is unable to exploit any relationship between the features and the labels, this is equivalent to the case where no such relationship exists [190, 191].

9.6.2 The bootstrap

The available sample **D** may be used to estimate the distribution of the items. This is done through the empirical distribution function, which assigns a probability mass of l^{-1} to the point in feature space occupied by each item. By sampling l items with replacement from **D**, one can generate many independent **bootstrap samples** of items generated from the approximated distribution. These many independent samples may be used to estimate the variability of a CV performance estimate derived from the sample. As l increases, the empirical distribution function will converge to the true one, making bootstrap inference asymptotically correct. Unfortunately, the empirical distribution can fail to represent the true distribution when l is limited, particularly in high dimensional contexts. One effect of this is that identical items will appear in the disjoint training and testing sets of a CV experiment on the bootstrap sample. This leads to a poor model for the distribution of $\bar{Q}$ in CV that includes a strong optimistic bias [159]. While the .632+ bootstrap [165] discussed in section 6.6.2 attempts to overcome the problem of bias (it is not widely recognised as having done so [36]), it is also not advanced by its creators as being directly usable for inference.

9.6.3 Remarks

There remain cases in ANA research where it is yet to be decided whether meaningful prediction is possible, and permutation testing may be used to provide evidence that this is so. However, this is not the case in the majority of AD ANA studies, which typically seek to improve on the work of many previous studies on a particular imaging feature. Here, it has already been shown that prediction is possible, and a study will instead wish to assess what performance is acheivable or how best to acheive it. Unfortunately, there is no way to extend permutation testing to provide pairwise tests or confidence intervals for learner performances, so it is not useful here.

Because the dimensionality of the feature space is typically very high relative to the number of items in AD ANA, typical samples are unlikely to provide a good enough approximation of the items' full distribution to allow for reliable bootstrap testing.

9.7 Discussion

In this section, I shall summarise the important themes that appear in the statistical procedures of the preceding sections. These include the problem of variance estimation for t-statistics in normal models, the issue of replicability discussed in the works of Bouckaert, the benefit of low variance CV strategies, and the use of calibration parameters.

9.7.1 All tests are heuristic

Apart from the permutation test, all the statistical procedures discussed in this chapter are inexact, as they are based on crucial simplifying assumptions that are not strictly correct. These might be independence assumptions in the variance estimation of t-like tests [37, 40, 178] or the neglect of any variability in a variance estimation [37, 167]. All methods assume that $\bar{Q}$ is normal. This can be justified by the central limit theorem under the fixed predictor model where all performance measures on the distinct items of $\mathbf{D}$ will be independent, but cannot be strictly justified in the general case. As discussed in 8.3.2, the true joint distribution of the performance results is unknown, and some assumption like this is inevitable.

9.7.2 Variance estimation

The bias of the variance estimator used in a t-statistic will determine the degree to which associated inference is conservative or permissive. Different assumptions will lead to different biases: assumptions of independence between experiment repetitions on a fixed dataset of the type that appear in the (uncorrected) resampled t test cause unacceptably high type I error rates and should be avoided. The assumption of independence between results on disjoint KCV test sets will result in a smaller bias that will normally be more acceptable.

In addition to bias, another important feature of a variance estimator is its distribution, which may be characterised with an effective DOF. Several of the approaches considered may underestimate the width of their variance estimator [37, 40, 167], though this should only be a significant problem when the true effective DOF is low.

9.7.3 Using the information from low variance CV strategies

Tests using multiple random partitions of a dataset can reduce the influence of internal randomness in the estimation of the mean performance and its variance [179]. This is a desirable feature of a procedures, as it has the potential to lift repeatability and lower error rates. Conversely, it is an unappealing feature of the conservative Z-test that it relies on internal randomness to avoid being underpowered.

9.7.4 Calibration parameters

Many of the procedures used calibration in synthetic problems to either estimate the width of their test statistic's distribution [178, 179] or find a compromise between additional information and independence assumption violation [37, 40]. This is unappealing, because it relies on the simulated problem being similar to the real problems to which the test is applied. While a conservative choice of value may seem appealing, this could lead to underpowered procedures.

9.8 Practice of inference in AD ANA

In this section, I shall discuss the role of inference in ANA, as well as the common practices in ANA for AD. I shall highlight some deficiencies in current practice and make suggestions for how it may be improved.

9.8.1 Common practice

A review of the AD ANA literature shows that significance testing is occasionally used for the following two tasks:

- pairwise comparison of methods to demonstrate the superiority of a method to a reference [28, 50, 128] and
- tests against the null hypothesis that a learner performs no better than chance [88, 96].

With a few exceptions [53], confidence intervals for performance quantities are rare.

The majority of studies using KCV or related CV strategies report no statistical analysis at all [29] (examples in [192–194]). Most commonly, methods are compared on the basis of point estimated performance alone.

Studies using SHOCV more commonly include statistical analyses [25, 28, 88, 96], though this analysis must be most strictly interpreted as making statements about predictor performances, rather than learner performances.

Various studies treat the results on the separate test sets of KCV as independent, and use this to conduct significance tests and produce confidence intervals [50, 128, 142, 195]. A few studies make the mistake of treating the results of sequential experiments using different partitions of a single dataset as if they were independent, thus producing tests with type I error rates that are potentially 1.0 (see section 8.5.3). These include several examples of a procedure similar to the resampled t-test [45, 102, 115], where the component train-test results of RHOCV are assumed to be independent, and several instances where the sequential KCV experiments in RKCV are assumed to be independent [47, 50].

In the cases where authors wish to test against the null hypothesis that a learner performs no better than chance on the basis of KCV results they have done this with the permutation test [137, 190]. None of the other specialist procedures discussed in this chapter are used.

9.8.2 Discussion and recommendations

While it is important that analyses do not lead to large numbers of false positives, it is my belief that statistical practice in AD ANA has been negatively affected by undue concern about the strict validity of inference procedures. I suspect that it is these concerns that are responsible

for the typical absence of statistical analysis. This is regrettable, as even a somewhat flawed statistical analysis is certainly preferable to none. Until new procedures with better validity guarantees are established, one of the following inexact procedures should be used.

- The cross validated t-test in which the test set results of KCV are assumed to be independent. This will be more reliable when K takes a lower value, as lower values of K are normally associated with lower values of C.
- For classification problems, a binomial model test in which the item performance results of KCV are assumed to be independent. While the greater correlation between results in the same test set may itself lead to more permissivity, the use of a binomial model over a normal one should avoid the problems discussed in 8.2.2.2.
- The corrected resampled t-test. This should allow inference to be performed with a low variance CV strategy where this is computationally feasible. As previously discussed, the test could be improved by limiting the DOF to $l-1$.

None of these procedures are able to account for dependencies not expected under the fixed predictor model, and they may all be moderately permissive. However, none of them rely on awkward calibration parameters, and none of them use deliberately noisy test statistics that imply limited effectiveness. They are all relatively straightforward, and do not place heavy restrictions on the m that can be used. Crucially, none of them neglect the correlations caused by the reuse of items in testing. Procedures that do this (see [45, 46]) should be avoided.

While the use of SHOCV may seem to offer a solution to the problem of dependency, this is something of an illusion. As discussed in section 8.4, the use of fixed predictor models in that context will provide statistical statements that are only strictly valid for the performance of the specific predictor constructed in that experiment, rather than that of the learner that selected it. The use of SHOCV is also associated with undesirable high variance performance estimates that will be associated with relatively low power. Where computationally feasible, inference should be conducted with other strategies instead.

By far the most common statistical analysis in AD ANA is a pairwise comparison of learners to demonstrate improvement, with interval estimation being relatively rare. Confidence intervals should be used where possible, as they are more informative when conducting a comparison of methods across different studies [174]. This can be an addition to any comparative tests.

In the cases where the demonstration of better than chance performance is of scientific interest, the permutation test should be used.

Summary

I have reviewed specialist inference approaches to CV inference from the literature and identified their key strengths and weaknesses. Specifically, I find that the use of many train-test repetitions should be beneficial, as it promotes the greater replicability required for an effective procedure. Calibration based on a single problem cannot be trusted to provide the right balance between power and type I error rates, and the results from sequential experiments using the same items for testing must not be modelled as independent. I have reviewed the practice of performance inference in ANA. This is quite sparse, as the majority of studies reach their conclusions on the basis of point estimation alone. More statistical treatments should be used to aid in the interpretation of experiment results. Confidence intervals in particular should be reported more often. While researchers may be concerned that statistical treatments may not be strictly valid, they should note that even a flawed analysis is more informative than none. There are several candidate procedures from the literature that should offer acceptable power and error rates.

Chapter 10

Extended inference procedures

This chapter is devoted to the development of new specialist statistical procedures for inference in cross validation (CV). In it, I shall describe the motivation for such procedures and consider two potential approaches: one based on the median combination of p values, which I call voting, and one based on the combination of test statistics, which I call bolstering. I shall then validate the bolstering approach in two classification problems: one real and one synthetic.

The bolstering approach developed and validated here was presented at the International Workshop on Pattern Recognition in Neuroimaging in 2015 [196].

10.1 Motivation

As discussed in section 9.8.2, statistical inference in AD ANA is patchy and flawed. Conventional analyses based on fixed predictor models may fail because of the problem of dependency, so alternative methods may be preferred. Unfortunately, the alternative heuristic approaches described in chapter 9 suffer from a number of undesirable characteristics. These including the following:

1. potentially permissive behaviour even under fixed predictor models,
2. the use of potentially unreliable calibration parameters (see section 9.7.4),
3. an incompatibility with high repetition, low variance CV strategies (see section 9.7.3),
4. an incompatibility with RKCV and other more efficient CV strategies (see section 9.7.3),
5. an incompatibility with binomial models that may be necessary in classification problems (see section 8.2.2.2).

If new heuristic statistical procedures can be developed that overcome these limitations, it should be possible to achieve greater power and lower error rates. If it can be shown that the new procedures are reliable, this should increase the use of statistical inference in AD ANA

and related applications. This in turn will make published results more interpretable, leading to better direction of future research efforts and a diminished risk that research methods will be put into practice on the basis of spurious benefits.

10.2 Conservative extension rules

In my development of new statistical procedures, I have taken inference procedures based on fixed predictor models as 'baseline' approaches on which to improve. These baseline approaches use a single 'base' repetition of a CV experiment based on a random partition of the available dataset (i.e., KCV or SHOCV). Improvement is achieved by incorporating the additional information provided by additional base experiment repeats using different randomly generated partitions. By guaranteeing improvements over the fixed predictor model inference, particularly in type I error rates, I hope to offer inference procedures that can be seen as sufficiently conservative to be used in AD ANA.

The joint distribution of the performance results associated with the sequential base experiment is unknown, so it is difficult to know the best way to combine them to produce a useful test statistic. Rather than using calibration, I have tried to produce **conservative combination rules** that try to exploit the additional information afforded by base experiment repeats without overestimating how much this is. These rules should produce inference that is consistent with the fixed predictor model. Even when no additional information is offered by additional experiment repetitions, the use of multiple repeats instead of a single one should not lead to more permissive inference. This should be true when the base CV experiment is either KCV or SHOCV.

Without assuming knowledge of the level of dependency between the results of base CV experiment repetitions, it is not possible to produce exact p values. This means that the actual type I error rates of the extended procedures are unknown. It also means that is not possible to use the additional information provided by sequential repetitions to guarantee an increase in power at all effect sizes. Instead, the benefit of the extended inference is guaranteed in terms of increased *repeatability* over the baseline procedure. In practice, this should lead to lower type I error rates and greater power to detect large effects.

10.2.1 Repeatability as an objective

Consider two tests denoted $T1$ and $T2$. For all effect sizes, the most likely outcome (e.g. reject/do not reject H_0) is the same, but $T2$ has greater repeatability than $T1$. There $T1$ has power less than 0.5, $T2$ has yet lower power. This will occur under the null hypothesis, meaning that $T2$ will have lower type I error rates. At large effect sizes where $T1$ has a power > 0.5, $T2$

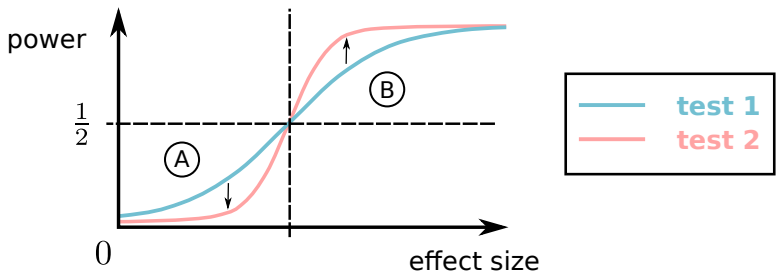

Figure 10.1: Power curves of statistical tests. Test 2 is an ideal extension of test 1. In effect size region A, which contains the null hypothesis, extension reduces the power of test. Note that the intercepts of the power curves correspond to the type I errors of the corresponding statistical test. In region B, extension increases power.

has a greater power. This is illustrated in figure 10.1.

I advance that $T2$ is to be considered superior to $T1$, and thus repeatability can be used as a metric to compare statistical tests.

In the context of confidence interval construction, increased repeatability will mean that points near the true value of the parameter to be inferred are included in an interval more frequently, while points further away from it will be included less frequently.

10.2.2 Paradigms of use

I intend the extended inference procedures to be used in two ways:

- A powerful paradigm based on RKCV, with KCV as the base experiment. This should provide non-permissive inference that is more repeatable than the cross validate t-test taken as a baseline and more efficient than the corrected resampled t-test (based on RHOCV).

- An alternative conservative paradigm based on RHOCV or EKCV, with SHOCV as the base experiment. Inference based on SHOCV may be used in AD ANA over inference based on KCV due to concern that the latter is more susceptible to problems of dependency. The effects of extension in RHOCV/EKCV should be more drastic than in RKCV, as extra repeats will mean more items are used for testing than in the baseline test. Extended inference in RHOCV/EKCV using SHOCV as the base experiment should be even more conservative than SHOCV-based inference, and should provide greater power.

10.3 Voting or median p value combination

In this section I shall consider the voting method of extending a baseline inference procedure for a single CV experiment to work with many CV repetitions on the same dataset. To do this, the baseline inference procedure is conducted using the results of each of the E repetitions

individually to produce a series of p values denoted $\langle p_e \rangle_{e=1}^{E}$, where E is assumed to be odd. The outcome of the voting test is decided with by taking the median p value. This is equivalent to combining the reject/do not reject decisions of the E experiments by majority vote.

The voting procedure is part of a broader family of p value combination methods that use some rule to produce a combined p value p_* from the sequence $\langle p_e \rangle_{e=1}^{E}$. To use any of these methods to construct confidence intervals, one must shift the location of a boundary until it corresponds to a null hypothesis producing the appropriate value of p_*.

10.3.1 Previous approaches based on the combination of p values

This section will describe previous method for combining p values and describe why they are not appropriate. This will provide context and motivation for the voting rule.

10.3.1.1 Approaches with independence assumptions

There are several long established methods for the combination of independent p values. The oldest of these is Tippet's method,

$$p_* = 1 - (1 - p_{\min})^E \text{, where } p_{\min} = \min p_e \,, \tag{10.1}$$

which assumes only that the p_e are i.i.d uniformly distributed on the interval $[0,1]$ under the null hypothesis . There is also Fisher's method, which uses the same assumption and derives p_* from the test statistic

$$-2 \sum_{e=1}^{E} \ln(p_e) \,, \tag{10.2}$$

which is modelled with a χ^2 distribution with $2E$ degrees of freedom. Neither of these methods is appropriate for combining the component results of repeated CV experiments on a shared dataset, as both assume that the p_e are independent.

10.3.1.2 Approaches with normal test statistic models

In addition to the two classic methods, there are two others that do not assume the p_e are independent. In Brown's method, the p_e are assumed to be derived from normally distributed test statistics with a known covariance [197]. This is generalised in Kost's method [198], where the covariance structure need only be known up to a scalar multiplicative constant. Unfortunately, the covariance structure is unknown in repeated CV experiments, so these methods are not appropriate either.

10.3.1.3 Approaches without assumptions

If the individual p_e are drawn from continuous statistics and provide valid inference under the null hypothesis, then it can be assumed that each p_e is marginally uniformly distributed on the

interval $[0,1]$. Any multivariate distribution for which the marginal distribution of each variable is uniform may be called a copula [199]. The joint distribution of the p_e is then a copula of $[0,1]^E$. The cumulative probability of a copula must lie between two limits termed the Fréchet-Hoeffding bounds. In the case the p_e, this result can be used to show that the p value

$$p_* = \frac{2}{E}\sum_{e=1}^{E} p_e \tag{10.3}$$

will lead to strictly non-permissive inference [200], regardless of the precise joint distribution of the p_e.

Though this combination method should produce conservative inference, it risks very low power: the distribution of $p_*/2$ will tend to be narrower than that of p_e about the same expectation (see appendix E). Because p values are bounded, even when $p < \alpha$ with some considerable probability, $\mathbb{E}[p]$ may be well above α, and $2\mathbb{E}[p]$ will be far above α. Thus, even when a baseline test with a single p_e has considerable power, a test with p_* may have low power.

10.3.2 Analysis of the voting rule

Recall the discussion of replicability and repeatability in section 8.6 of chapter 9. The outcome of a test in the eth CV experiment is the random variable $Y_e \in \{0,1\}$, with 1 signalling null hypothesis rejection. A given random sample specifies the random variable $W \in [0,1]$, which is the probability of selecting a CV design that leads to a positive outcome on that test. By definition, $P(Y_e = 1|W = w) = w$, and $P(Y_e = 1) = \mathbb{E}[W]$.

10.3.2.1 Replicability

The proof of appendix D shows that replicability must increase when an odd number of experiment repetitions E is increased by 2 in a voted test. This shows that the extended test with odd $E > 1$ will have greater replicability that the baseline test. This is necessary for greater repeatability, but not sufficient.

10.3.2.2 High E limit

It is instructive to consider the behaviour of a voted test in limit where $E \to \infty$, as the finite E test will approach this behaviour as E is increased. In this limit, majority vote will lead to a positive outcome *if and only if* $W > 1/2$. Because W is bounded in $[0,1]$, the following Markov inequalities hold

$$P(W \geq 1/2) \leq 2\mathbb{E}[W] \text{ and} \tag{10.4}$$

$$P(W \leq 1/2) \leq 2\big(1 - \mathbb{E}[W]\big). \tag{10.5}$$

Note that $\mathbb{E}[W]$ is the marginal probability of a positive outcome in a test using the baseline test based on single experiment repetition. By inequality (10.4), when the type I error rate of the baseline test is less than 0.5, the type I error of the extended test is no more than twice that. A consequence of this is that a test based on a p_* defined as twice the median p value will be strictly non-permissive. (This echoes the result for the mean p value in 10.3.1.3). By inequality (10.5), where the type II of the baseline test is less than 0.5, the power of the extended test is no more than twice that.

Stronger proofs are possible if there is an assumed distribution for W. The proof in appendix C shows that, if the distribution of W is unimodal and symmetric,

$$P(W > 1/2) > \mathbb{E}[W] \text{ when } \mathbb{E}[W] > 1/2, \text{and} \tag{10.6}$$

$$P(W > 1/2) < \mathbb{E}[W] \text{ when } \mathbb{E}[W] < 1/2. \tag{10.7}$$

This guarantees an increase in repeatability for the extended test when infinitely many repeats are used. It can be demonstrated numerically that the same inequalities hold true when W has a Beta distribution.

10.4 Bolstering

In this section, I shall describe an extension rule based on the mean combination of test statistics from CV repetitions. I call this rule **bolstering**. The name is derived from the word 'bolster', which means 'to strengthen'.

The essential idea is to take a statistic Ξ that is exactly or approximately (multivariate) normally distributed under a fixed predictor model for KCV. The statistics $\langle \Xi_e \rangle_{e=1}^{E}$ from the E separate KCV repetitions are combined together to form a **bolstered statistic**

$$\bar{\Xi} = \frac{1}{E} \sum_{e=1}^{E} \Xi_e \tag{10.8}$$

which is modelled as if it has the same distribution expected of Ξ_e in the baseline test. Inference based on the bolstered statistic may be called **bolstered inference**.

10.4.1 Analysis of bolstering rule

10.4.1.1 Narrowing of test statistic distributions

The distribution of $\bar{\Xi}$ will in general be narrower than that of a Ξ_e. As shown in the proof of appendix section E, when taken as an estimator of $\mathbb{E}[\Xi_e]$, $\bar{\Xi}$ will have an expected error that is

less than or equal to that of a Ξ_e for all errors defined

$$\mathbb{E}\left[\,|\Xi - \mu|^z\,\right] \text{ for } z > 1\,. \tag{10.9}$$

A consequence of this is that $\mathbb{V}\text{ar}[\bar{\Xi}] \leq \mathbb{V}\text{ar}[X_e]$, with equality holding only when the Ξ_e are perfectly correlated and there is no variability associated with the random selection of CV partition. Where Ξ is vector valued, $\mathbb{V}\text{ar}[\bar{\Xi}\cdot x] \leq \mathbb{V}\text{ar}[\Xi\cdot x]$ for all x.

10.4.1.2 Increased repeatability

In the univariate case, whenever $\bar{\Xi}$ and Ξ_e share a symmetric unimodal distribution form (e.g., they are both normally distributed), this will mean that

$$P(\bar{\Xi} \geq \xi) \geq P(\Xi_e \geq \xi) \text{ when } \xi < \mu,\ P(\Xi_e \geq \xi) > 1/2\,, \text{ and}$$
$$P(\bar{\Xi} \geq \xi) \leq P(\Xi_e \geq \xi) \text{ when } \xi < \mu,\ P(\Xi_e \geq \xi) < 1/2\,.$$

Where the sign of $\Xi - \xi$ decides the outcome of a significance test (by determining a p value) this demonstrates that bolstered test will have a greater repeatability than the baseline test. This scheme illustrated in figure 10.2.

The assumption of normality for the test statistic is not easy to justify, but it is one shared by almost all statistical procedures. The normality (or approximate normality) of the Ξ_e will often be expected in the fixed predictor model. For instance, where $\bar{Q}$ denotes the mean performance observed in KCV, and S^2 represents one of the variance estimators discussed in 9.3, both $\bar{Q} - \mu$ and $\bar{Q}/S$ will be approximately normal.

In the rare case where m/l is small, if E is very large, one could justify the normality assumption for the mean test statistic taken over many repetitions by its definition as an approximate U-statistic(see 9.5).

In the case of multivariate test statistics, when the decision boundary is linear, or close to linear, one can consider only the projection $\Xi\cdot x$ of tests statistic onto the normal x of the decision surface. This reduces the problem to a univariate one, so repeatability should increase as before. Because decision boundaries are smooth, linear approximations will work well at smaller scales. An illustration of the decision boundary associated with the sample mean and variance in a one-sided t test with 4 degrees of freedom is provided in figure 10.3.

10.4.1.3 View of opposing forces

The correlations between the component results of the base CV experiment are likely to make the distribution of that test statistic wider than expected under the fixed predictor model, mak-

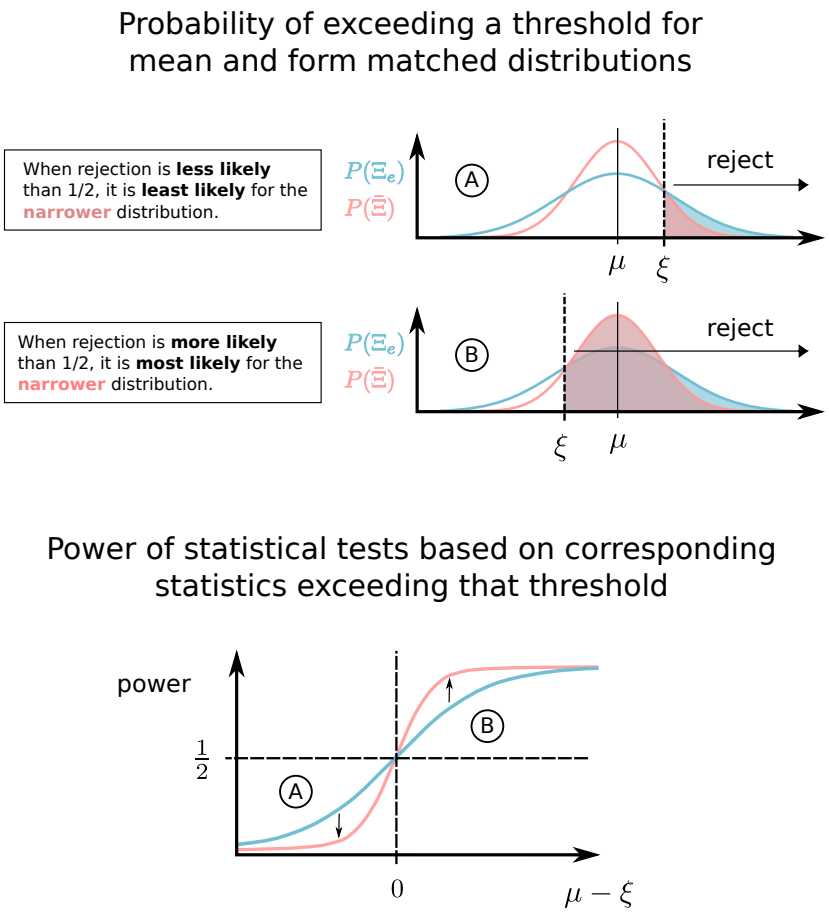

Figure 10.2: The use of a bolstered statistic increases the repeatability of statistical tests.

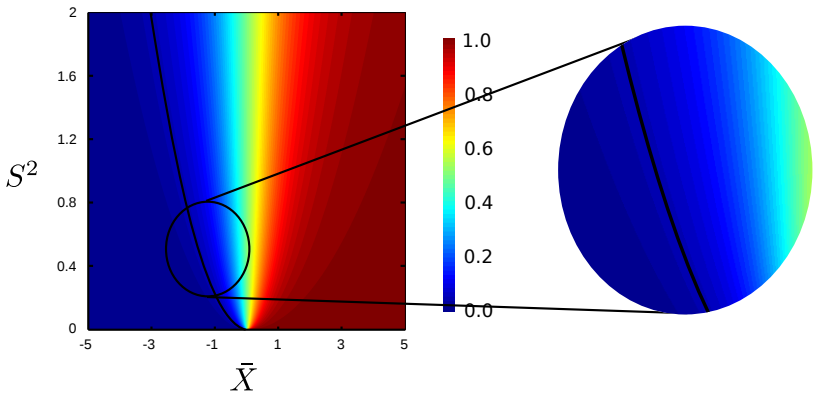

Figure 10.3: Decision boundary in a one-sided t-test. S^2 represents the sample variance estimate, and $\bar{X}$ represents the sample mean.

ing the resulting inference permissive. The 'smoothing' over many repetitions will make the distribution narrower, making the resulting inference conservative. This scheme is illustrated in figure 10.4. The precise balance of these opposing forces is difficult to predict in advance, so bolstered inference may be either conservative or permissive.

Though the two forces have opposing effects, they have the same cause: the variability of predictors. I conjecture that their occurrences in learning problems will be highly correlated. If this is true, then bolstering should offer the most improvement (test statistic stabilisation) in the cases where it is needed the most (where the problem of dependency is important).

10.4.2 Implementation of bolstered procedures

The specification of a test statistic in a base test may be ambiguous, but this choice will affect how the bolstering combination rule is applied. For instance, where some measurement X_e is taken in a base experiment, either X_e or X_e^2 might be used as a test statistic, but $\bar{X}^2$ may not be equal to $\bar{X^2}$.

Care must be taken when applying the bolstering rule to extend a baseline test, as a poor choice of the statistic to be average will lead to poorer test behaviour. A good choice of statistical will be one that can reasonably be expected to have an approximately symmetric unimodal distribution. It may also be desirable if the statistic would be an unbiased estimator of some population parameter under the fixed predictor model (e.g., learner performance, variance), as this property will be retained by the bolstered statistic. This means that the same statistic may be used for inference and interpretation.

This section presents two implementations of the bolstered statistical procedures: one for the binomial model, and one for the normal model.

10.4.2.1 Bolstered binomial inference

As described in section 8.2.2, under the binomial model appropriate for classification, the only required test statistic in inference for the performance of single learner is the mean performance $\bar{Q}$. The only required test statistic in the comparison of two learners is the 2×2 contingency table.

In order to produced bolstered procedures for binomial inference, I begin with a model that assumes that all item performance measures in the base experiment are i.i.d. As discussed in section 9.3, in KCV, this type of assumption may lead to a greater degree of bias in the estimation of the variability of $\bar{Q}_e$ than an assumption of independence between test set results. However, this assumption is necessary to retain the binomial model, which itself is important in preventing permissiveness. (As discussed in 8.2.2.2, a normal approximation for binomial per-

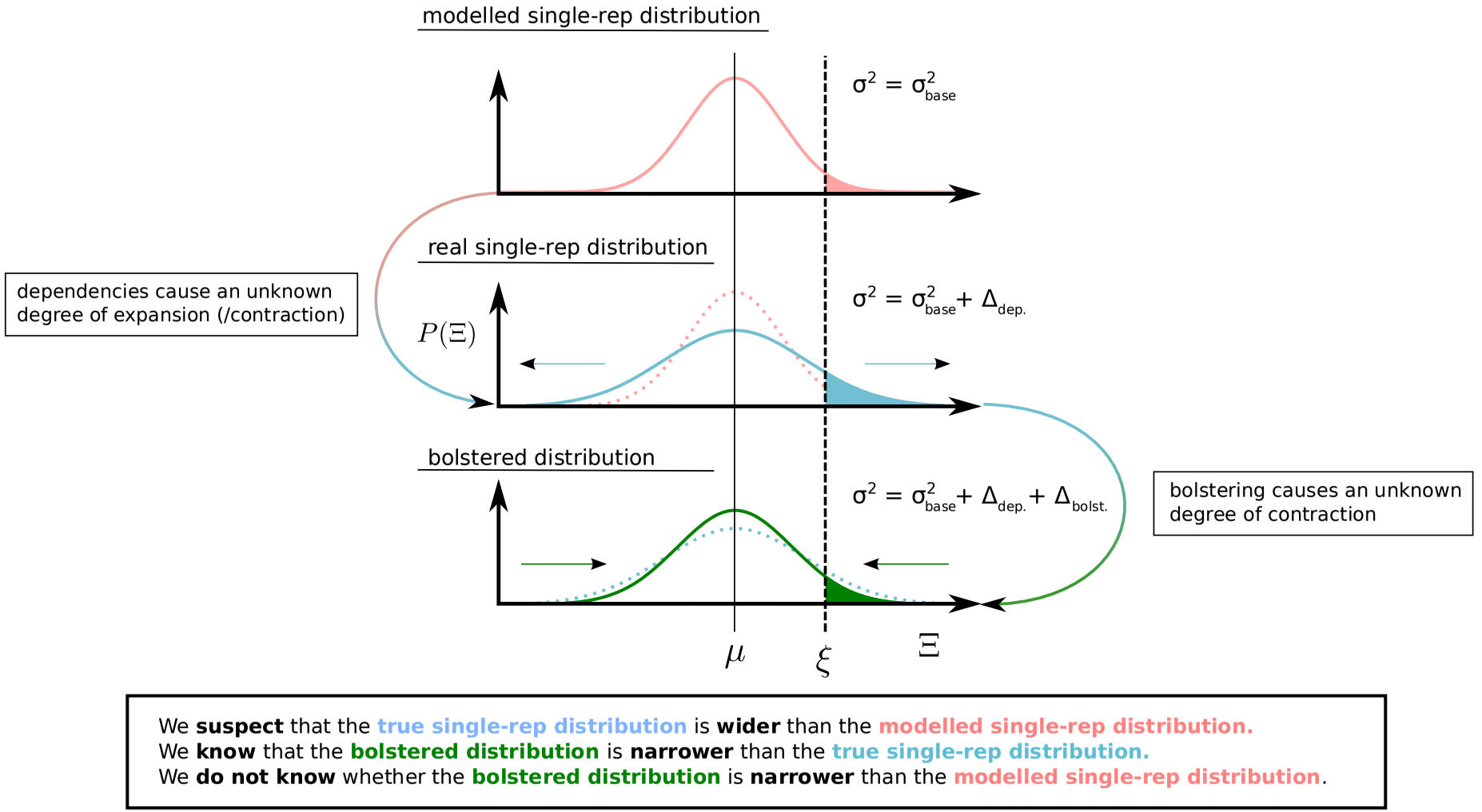

Figure 10.4: A trade-off between two unknowns in determining the variance of the bolstered statistic. The effects of dependencies and smoothing over multiple repetitions are modelled with variance contributions.

formance measures can lead to permissive inference even under fixed predictor assumptions.)

In inference for a single learner performance, one takes the average performance $\bar{Q} = \sum_e \bar{Q}_e$ over all base CV repetitions. Where n' denotes the number of items used for testing in the base experiment (n in RHOCV and l in RKCV), the quantity $n'\bar{Q}$ is assumed to be binomially distributed n' trials. For pairwise comparison of learners, one takes the average of the 2×2 contingency tables to performance bolstered McNemar's test. It is sufficient to take the average of the quantities A_1 and A_2 as defined in section 8.2.2.1.

The binomial cumulative distribution function is computed using the regularised incomplete beta function. This provides a smooth analytic continuation of the function to non-integer values.

10.4.2.2 Bolstered normal model inference

To produce bolstered inference based on normal models, the base experiment test statistic should be the two element vector $(\bar{Q}_e, S_e^2)$, where $\bar{Q}_e$ represents a mean performance estimate, and S_e^2 represents an estimator of $\bar{Q}_e$'s variance. The bolstered normal model inference should use the t-statistic $\bar{Q}/S$, modelled with the same DOF expected in the base test.

When using RKCV, the variance estimator S_1^2 of section 9.3 should be used over S_2^2 to avoid the latter's potentially greater bias. This will mean that the bolstered inference is an extended version of the cross validated t-test, which can be expected to have greater repeatability.

The bolstered t-test for RKCV may be compared to the Bouckaert's calibrated tests [179]. Both use a normal model and combine the results of sequential KCV repetitions. However, the bolstered t test uses a variance estimator of limited bias, and does not require calibration to avoid grossly permissive behaviour.

10.5 Shared concerns

This section discusses two concerns relevant to both the voting and bolstering extension rules: the appropriate choices of the number of base experiments E, and the choice of K when RKCV is used.

10.5.1 Choice of number of base experiment repetitions

There is no theoretically optimal selection of E, the number of base experiment repetitions to be used in an extended test. E might seem like an undesirable free parameter, but it is not. This is because there are no values of E that will lead to inference behaviour that is worse than that of the baseline procedure. A choice of $E = 1$ simply implements the baseline procedure, and any increase on this should only improve repeatability.

The selection of E is then a trade-off between computational cost and repeatability benefit.

In practice, because the benefit of including further experiment repetitions will decrease with E, some reasonably large number (e.g., 100) should be sufficient to provide almost all of the possible benefits.

One could devise stopping rules to decide when E need not be increased further. Because the results of the sequential experiments are independent conditional on the dataset, one can use standard statistical analysis to for the convergence of the combined statistics. In the bolstering extension rule, one could wait until the one has estimated the conditional expectation of the Ξ_e to within some δ as measured by a confidence interval. In the voting extension rule, one could continue performing experiments until the latent rate parameter W was estimated to within a given precision.

Sequential analysis could be used to devise stopping rules for E with guarantees on replicability [201].

10.5.2 Appropriate choice of K in RKCV

Lower values of K should be used with the extended procedures. As discussed in part II of this thesis, these allow more accurate performance estimation at a given computational cost. Their greater variance reduction will also provide greater repeatability increases in extended inference procedures.

10.6 Comparison of the voting and bolstering extension rules

Though I think both of the extension rules could be used to improve inference in AD ANA, *I have focused my attention on the bolstering rule, and it is only this rule that I shall now validate empirically*. This is because the assumptions required to argue increased repeatability in the bolstering rule seem more plausible than those required for the voting rule; all other statistical procedures discussed in chapter 9 assume that various key test statistics will be approximately normal. I have little reason to assume that W has a distribution that is Beta or symmetric unimodal.

The bolstering rule also has the appealing feature that its final test statistics can be minimum variance estimators of population parameters. Under the binomial model, all that is required to produce a bolstered confidence interval for a learner performance is the mean observed performance used to estimate it.

10.7 Validating bolstered inference in a synthetic problem

This section presents an empirical demonstration of the behaviour of a bolstered test in the powerful paradigm based on RKCV with $K = 2$. It will be use a baseline test with a binomial

model. The use of a synthetic problem allows test powers, intervals coverage rates and learner performances to be measured to within arbitrary precision through experiment repetition on many independent samples.

The synthetic problem used in this study is the binary classification task first described in section 7.3.1.2. In this task, there are two classes of items that have d-dimensional Gaussian distributions separated by a distance of 2 in the first dimension. The distributions have identical covariance matrices with a variance of 1 along every dimension a covariance of ρ between all pairs of dimensions. Two example samples of items generated using $d = 2$ are presented in figure 10.5.

Figure 10.5: Different samples generated in the synthetic classification task with $d = 2$ using different values of the parameter ρ. Items of a given class share the same colour.

Two learners are evaluated. The first of these is linear discriminant analysis (LDA) (see section 2.6.1.3) which attempts to estimate and exploit the covariance structures, and nearest centroid (NC), a simple method which makes no use of the covariance structure. When the parameter ρ is low, NC has higher accuracy than LDA, as there is no covariance structure to exploit, and LDA's attempt to do so causes it to make errors. As ρ increases, the class distributions become more separated, and LDA is able to exploit this change to achieve higher performance. At sufficiently high ρ, LDA has a greater learner performance than NC. The change in the learner performances of LDA and NC with ρ is illustrated for training sets containing 60 items of each class in figure 10.6.

10.7.1 Description of experiments

In this study, I arbitrarily chose $d = 12$, and varied ρ between 0 and 1 in steps of 0.01. There were 6 different experimental settings corresponding to different choices for E, the numbers of KCV repetitions to be used in the bolstered inference. These were the elements of the set $\{1,2,4,8,16,32\}$. For each value of ρ and E, I generated 10^6 independent samples comprising 60 items of each class. On each of these I use the bolstered binomial inference procedures described in section 10.4.2.1 to generate confidence intervals for the performance of both learners and perform a pairwise comparison.

The intervals were two-sided and based on the Agresti-Coull procedure [177]. Their nom-

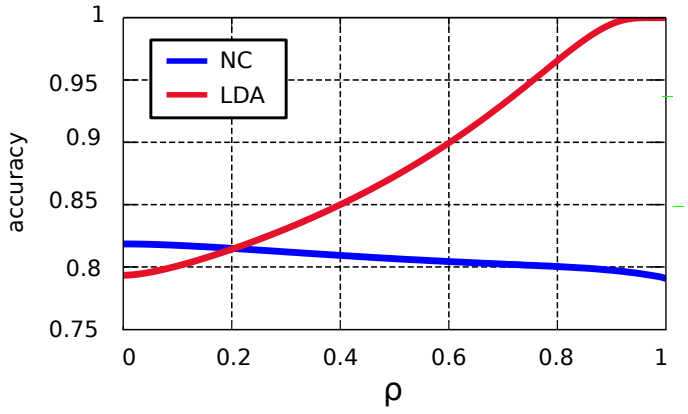

Figure 10.6: LDA and NC learner performances with training sets comprising 60 items of each class as ρ is varied.

inal coverage was set to 95%. The comparison was a one-sided McNemar's test against the null hypothesis that the performance of LDA was no more than that of NC.

The true learner performance associated with each value of ρ was taken as the average performance measured in the 10^6 independent repetitions of RKCV with $E = 32$. The coverage of an interval procedure was measured as the fraction of its intervals to contain the true learner performance, and the power of the McNemar's test was measured as the fraction of repetitions in which the null hypothesis was rejected.

I have written an efficient `C++` to conduct various CV experiments of the type described quickly. The experiments described in this section were run on approximately 50 compute nodes of a grid engine within a single day.

10.7.2 Results

This section describes the results of the experiments on the synthetic classification problem.

10.7.2.1 Accuracy distributions.

Figure 10.7 presents area normalised histograms describing the observed distribution the mean performance estimate $\bar{Q}$ for both learners at the arbitrarily chosen ρ value of 0.63. The distributions associated with $E = 32$ and $E = 1$ may be compared with the distribution expected under a fixed predictor model.

It can be seen that the distribution of the average performance $\bar{Q}$ where $E = 1$ is much wider than expected under the fixed predictor model for the NC learner. For LDA, the distribution of $\bar{Q}$ in this case is only subtly wider than expected under the fixed predictor model.

For both LDA and NC, the use of 32 repetitions makes the distribution of $\bar{Q}$ appreciably narrower. For NC this makes the variance close to that expected under the fixed predictor model.

For LDA, the $E = 32$ distribution is even narrower than the fixed predictor model, suggesting that inference may be conservative in this case.

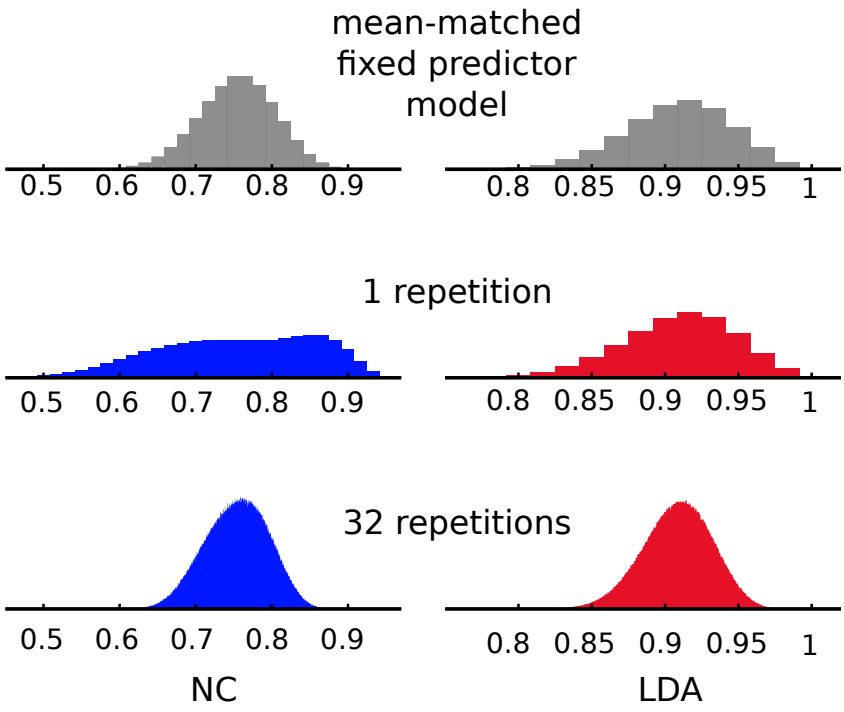

Figure 10.7: Area normalised histograms describing the observed distribution the mean performance estimate $\bar{Q}$ in the experiments of section 10.7. Results are presented for both learners at the arbitrarily chosen ρ value of 0.63. At the top of the figure are the distributions expected under a binomial fixed predictor model in which the predictor performances equal the learner performances, immediately below them are the distributions observed in RKCV with $E = 1$, and at the bottom are the distributions observed with $E = 32$. It can be seen that the higher value of E results in narrower distribution.

I believe that the greater deviation of the NC $E = 1$ distribution is due to the greater instability in the predictors selected by NC at this ρ value. This deviation appears to increase with ρ. A potential explanation for this is the following: as ρ increases, the variability of the centroids in a sample of items along the direction perpendicular to the axis that separates the two distribution increases. This causes a greater variability in the angle of the decision plane predictor produced by NC. This greater instability in predictor selection causes a greater deviation from the fixed predictor distribution.

10.7.2.2 Confidence intervals.

The coverages of nominally 95% confidence intervals for the learner performances of NC and LDA at different values of ρ are illustrated in figure 10.8. For both learners, the coverages associated with $E = 1$ are below those expected under the fixed predictor model (even if the

inexactness of confidence intervals for a binomial proportion is taken into account). The drop in coverage is particularly pronounced for NC. This is consistent the wider than modelled distributions of the type illustrated in figure 10.7 and my conjecture about the instability associated with high ρ.

The problem of reduced coverage is partially, completely, or more than compensated for by use of additional KCV repetitions. Coverage is uniformly improved with increasing E. This can be explained by the observed distributions of accuracies; for both learners and all settings, at all accuracies further than 0.017 from the mean, the probability of obtaining an accuracy measurement $\bar{Q}$ at least that far from the learner performance is highest with a single repetition. As the containment of the mean in an interval is a monotonic decreasing function of the distance, the coverage probability must increase for all intervals extending at least 0.017 in both directions. As might be expected, the improvement associated with doubling E diminishes rapidly with E. The improvement in coverage is greatest where the initial deficit in coverage is highest (NC, high ρ). This is consistent with the idea that bolstering will offer greater improvement when it is needed the most.

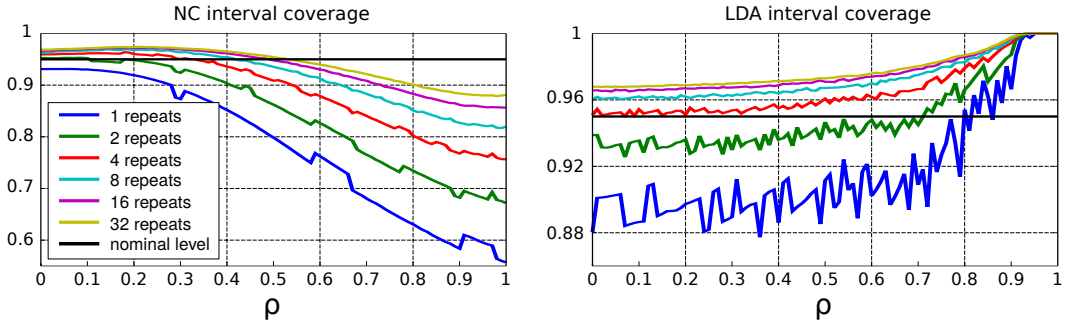

Figure 10.8: The coverage of confidence intervals for the learner performances in the synthetic problem at different values of the parameter ρ. The black line denotes the nominal coverage of 95%.

As illustrated in figure 10.9, the average width of confidence intervals was marginally increased by the use of additional repeats. The increase was greatest for NC at high values of ρ, where the use of 32 repeats of 1 increased the expected with of a confidence interval by approximately 4%. This increase in width seems a trivial cost to pay for the coverage improvement achieved in that situation.

10.7.2.3 Pairwise comparisons of the learner performances.

Power curves describing the power and type I error rates of a one-sided McNemar's test against the null hypothesis that LDA had a performance no higher than NC are provided in figure 10.10. A nominal significance level of $\alpha = 0.05$ was used. These are technically *pseudo*-power curves, as these are results specific to this problem. Here, the baseline test ($E = 1$) is not permissive,

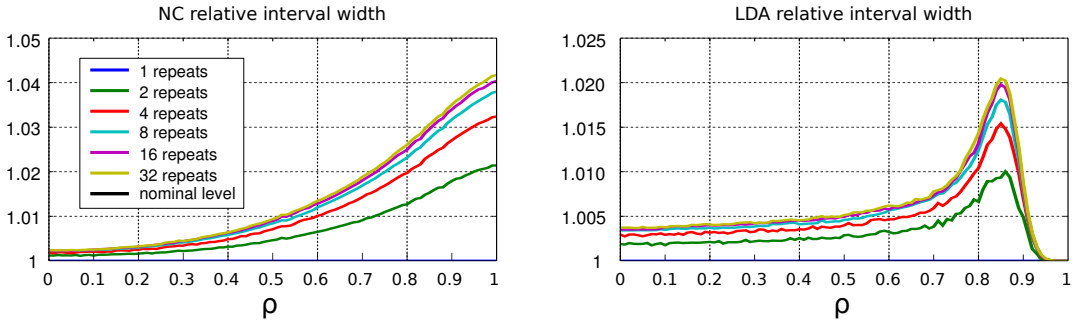

Figure 10.9: The average width of confidence intervals for the learner performances in the synthetic problem at different values of the parameter ρ. Width is presented relative to the width of the interval procedure using a single base CV repetition.

though the use of the multiple repetitions still improves the power curve from the repeatability perspective of section 10.2. As anticipated the power curves for the different E cross near 0.5.

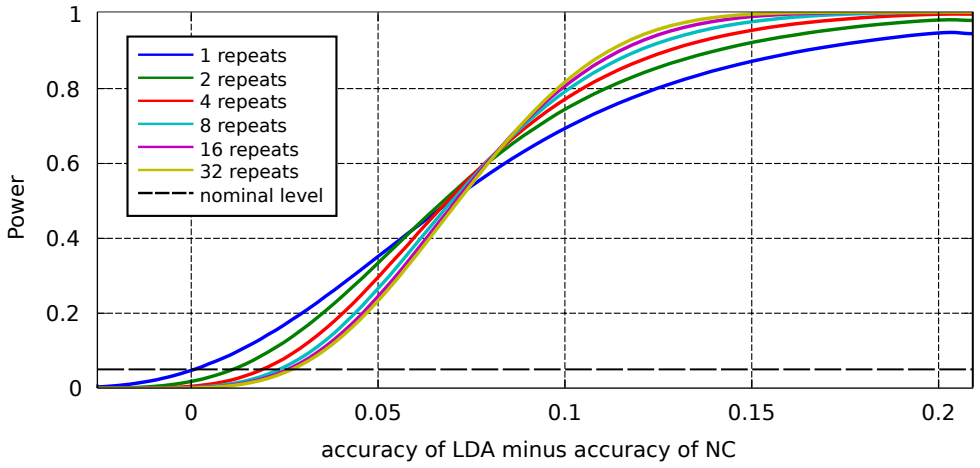

Figure 10.10: Power of the bolstered McNemar's test for pairwise differences in learner performances in the synthetic problem using the results of RKCV. This was a test against the null hypothesis that LDA performed no better than NC with a nominal α of 0.05.

10.7.2.4 Replication with RHOCV

I replicated the experiments of this section using RHOCV to implement bolstered procedures in the conservative paradigm. The results of the replication, presented in appendix F, are qualitatively very similar to those presented here.

There are two differences. Firstly, the problem of dependency appears to have stronger effect on the coverages of the intervals for the NC learner performance. Secondly, the effect of bolstering is much more pronounced; 32 repetitions is sufficient to make intervals for the NC learner performance conservative for all values of ρ, and the changes in the power of McNemar's test are roughly twice those seen in RKCV.

10.8 Validating bolstered inference in Alzheimer's disease classification

This section presents a study on bolstered inference in the conservative paradigm on a classification problem in the ADNI dataset. This RHOCV-based inference is intended as a superior alternative to the SHOCV-based inference used currently [25, 28, 88, 96].

The learning problem selected here was intended to have similar characteristics to the real problems studied in the AD ANA literature. For the sake of timely computation and implementation, I have used simple methods with relatively few feature selection and processing steps rather than the more complex methods that might be considered state-of-the-art. (I note that even if one of such methods had been used, there would be no guarantee that its behaviour would be any more representative of the wider set of methods seen in the field.) Within this constraint, I have aimed for a loose correspondence with the comparative study conducted by Cuingnet et al. [88], which used the SHOCV-based inference on which I would like to improve.

Because real data does not permit one to generate large numbers of independent datasets, it is not possible to measure interval coverages, test powers, and learner performances to arbitrary precision. For this reason, this study only reports estimators of the probability of decision *conditional on the dataset*. The consistency of results using different random partitions of a given dataset (i.e. replicability) will be greater than the consistency of results in independent datasets (i.e., repeatability). The replicability estimated by results in a single dataset, such as those presented here, indicates a limit on the repeatability , and thus also on power and type I error rates.

10.8.1 Description of the problem

This study considers the discrimination of AD subjects from healthy controls (HC) based on structural magnetic resonance imaging (sMRI). The binary classification task is one of the most well studied learning problems in ANA, and sMRI is the most commonly used imaging modality [24, 27, 29].

Aiming for correspondence with [88], I used 1.5 T T1-weighted magnetic resonance images from the baseline time-point of the ADNI study. To increase the available sample size (and so improve generalisation), I combined images from all phases of the study to obtain a full sample of 194 control and 171 AD subjects. Subjects with alternative possible diagnoses were excluded.

All subjects were registered to a groupwise template created with iterative non-linear registration. SPM12 software was used to produce grey matter maps on the images before registra-

tion, and these maps were modulated by the Jacobian determinant after they were transformed to the groupwise template. As was the case for most of the methods in [88], the final choice of features was the Jacobian modulated grey matter concentrations without spatial smoothing. Features were scaled to ensure that the median inter-point distance in the full sample was equal to one. A detailed description of the relevant methods is presented in chapter 2.

In order to examine the relationship between effect size and power in pairwise comparison tests, I needed to create some smoothly varying difference in performance between two learners. I achieved this by using the NC and support vector machine (SVM) learners, and smoothly varying the C parameter of the SVM. As discussed in section 2.6.1.1, this parameter controls the trade-off between accurate classification in the training sample and width of the separating hyperplane. Varying C produced the difference in performance presented in figure 10.11; at low values of C, the SVM learner performs poorly, as all items are assigned to the most common class. As C increases, the performance of the SVM learner improves to become better than that of the NC learner.

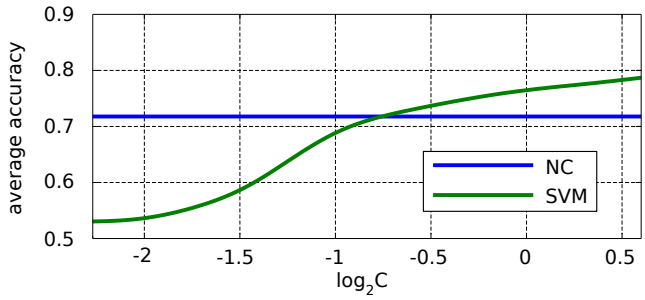

Figure 10.11: Average performance observed in all RHOCV experiments at different values of the parameter C in the AD classification problem. This figure illustrates how the parameter controls the expected difference in performance between the two learners.

10.8.2 Description of experiments

The value of the C parameter was varied over a range of 100 values spaced equally in the log domain between $2^{-2.27}$ and $2^{0.6}$. I chose this range based on preliminary experiments in which higher or lower values had little effect on the expected SVM performance.

For each value of the parameter C, I conducted 2×10^4 RHOCV experiments for all E values in the set $\{1,2,4,8,16,32\}$. As in [88], half the available items were used for training in each SHOCV experiment, and class stratification was used to preserve the balance between AD subjects and healthy controls. In each experiment, I produced 95% confidence intervals for the NC and SVM learner performances based on the bolstered Agresti-Coull procedure, and conducted a bolstered two-sided McNemar's test for the difference in performance between the

two learners with a significance level of $\alpha = 0.1$.

For each C value and each value of E, I recorded the fraction of times each potential learner performance value was included in the confidence intervals and the fraction of times the significance test indicated a difference in between the two learner performances.

Because of a then-unsolved bug related to compilation of the SVM on the grid engine, the experiments were performed on a single desktop machine. This and the relatively large sizes were the reason that the number of experiment repetitions used was much lower than that used in the study of section 10.7. The experiments were completed on a single machine over a long weekend.

10.8.3 Interpretation of results

In any real problem, the number of available samples is limited. This means that learner performances, test powers, and type I error rates cannot be measured to arbitrary precision, as it is not possible to produce large numbers of independent experiment results. Instead, I have conducted a study on replicability examining the reproducibility of the results on a single dataset. Rather than produce interval coverages, I have produced *containment* rates, which represent the probability that a potential learner performance value is contained in an interval conditional on the dataset. Rather than produce power measures, I have produced *detection rates*, which are the rate at which an effect is detected conditional on the dataset.

10.8.4 Results

This section describes the results for the confidence interval and pairwise comparison test procedures in the ADNI classification task.

10.8.4.1 Confidence intervals for the learner performance.

This section presents results for the SVM learner with the highest possible value of the C parameter, as this is the learner setting most representative of real AD ANA research [34, 88]. Rather than coverage estimates, I am reporting the rate at which potential performance values were contained in the interval conditional on the data. These are illustrated in figure 10.12.

With a single SHOCV experiment, the results of the interval procedure are variable, with even the expected performance value (0.79) being excluded in a certain percentage of cases. It can be seen that additional repeats move the containment rate toward either 0 or 1, depending on which is closer. This indicates an improvement in replicability.

10.8.4.2 Pairwise comparisons of the learner performances.

Figure 10.13 shows the fraction of pairwise tests that detect a difference between the learners. This graph is not symmetric, as the difference in accuracy is not the sole determining factor.

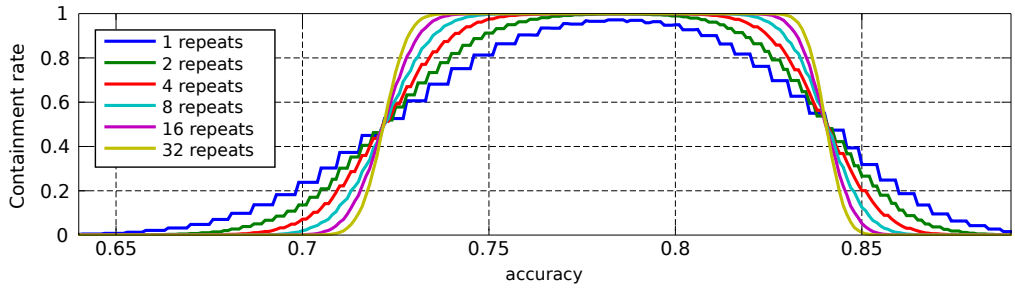

Figure 10.12: Illustrations of the fraction of intervals to contain possible values of the SVM learner performance at the highest value of C.

The replicability of the single-repeat test is very low; even where the estimated difference in accuracy is as high as 0.10, this is detected in less than two thirds of experiments. If the replicability in this sample is representative of replicability in the population, then this indicates that single repetition hold-out tests as used in [88] must also have even lower repeatability, and therefore have a power even lower than the detection rates observed here. (As well as higher type I error rates.) The detection rate is improved dramatically with further repetitions, and the curves cross near the halfway point, indicating improved replicability. This makes improved power plausible.

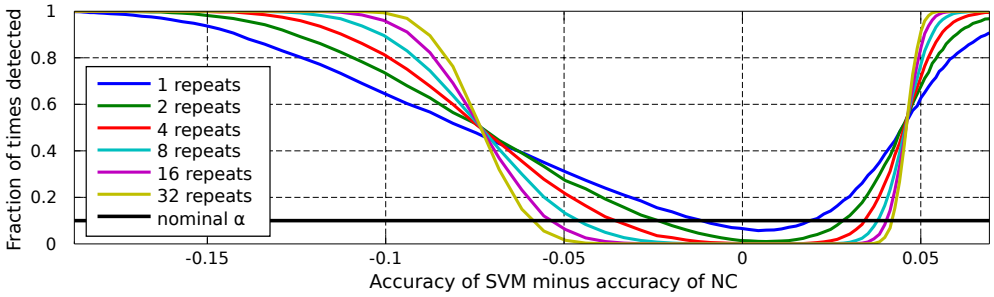

Figure 10.13: Detection rates of the bolstered McNemar's test for pairwise differences in learner accuracies associated with different values of the C parameter in SVM.

10.9 Discussion

The bolstered procedures for binomial inference provide clear benefits over the single repetition procedures that they are based on. Results in the synthetic experiment demonstrate increases in repeatability. This practically entailed improvements of the coverage of confidence intervals without significant increases in expected length, as well as improvements in the power and type I error rates of significance tests. Improvements were most pronounced in the cases where the problem of dependency was greatest, namely in interval procedures for the NC learner

performance at high values of the parameter ρ.

The amount of repeatability improvement associated with additional KCV repetitions diminished rapidly with the number used. This shows that the majority of the potential benefits can be achieved by some intermediate value of E (say 32) that is chosen in advance.

The results of the ADNI classification problem indicate that SHOCV-based inference in plausible AD ANA experiments can have low replicability, even at reasonable effect sizes. Detection rates, which are likely to be closer to the edges of the interval $[0, 1]$ than true power rates, were close to 0.5 for large regions when only a single SHOCV repetition was used. The use of more repeats moved power and detection rates towards the edges of the interval, providing the increased consistency necessary for higher power.

10.9.1 Use of bolstered inference in AD ANA

Because of their conservative construction, bolstered procedures can offer better repeatability than the baseline procedures based on single repetitions of KCV and SHOCV that are the only procedures used currently without potentially unlimited error rates (see section 9.8). The improvement in repeatability will practically mean higher interval coverages, lower test error rates, and higher power at large effect sizes. The use of bolstered procedures will also allow inference to be conducted simultaneously with low variance performance estimation.

In section 9.8.2, I called for more inference to be used in AD ANA. I specifically recommended that it be conducted with either the corrected resampled t-test, or a procedure based on a fixed predictor model in KCV. Because of the improvement in interval and test behaviour, *I suggest the bolstered procedures described in this chapter be used in AD ANA over their single repetition alternatives where computational resources allow this.* Because of their greater efficiency and compatibility with binomial models, I recommend that bolstered procedures based on RKCV be used over the corrected resampled t-test.

I primarily advocate the use of bolstered inference based on RKCV with a relatively low value of K, as this will allow for lower variance learner performance estimation that is more efficient than would be achieved with RHOCV. However, I anticipate concerns about the strict validity of the resulting inference. As can be seen in the result for the coverage of intervals for the performance of the NC learner in synthetic experiment (figure 10.8), bolstering procedures may still occasionally be permissive. If this is a crucial concern, procedures based on RHOCV may be used; as demonstrated by the replication of the synthetic study using RHOCV, the improvements in repeatability (and thus the reduction in type I error rates) will be more drastic. While it will still not be exact, bolstered inference based on RHOCV should provide a more conservative inference than any other strategy discussed in this thesis.

While bolstered statistical procedures are inexact, this property is shared by all statistical procedures for learner performance quantities based on CV (see section 9.7.1). Bolstering does not solve the problem of dependency. However, because the largest improvements in procedure behaviour are seen in cases where the problem of dependency is greatest, it can be viewed as a partial countermeasure.

10.9.2 Limitations of the validation study

Because of the dominance of classification tasks in AD ANA, the validation presented in this chapter has exclusively considered procedures based on binomial models. Ideally, the normal model procedures described in section 10.4.2.2 would be compared to the most viable alternatives described in chapter 9, such as the corrected resampled t-test. These models should be widely applicatible, as the central limit theorem provides a justification for their use in most cases.

Finally, the information provided by the study on ADNI data was limited by the experiment design. I note that a resampling design based on disjoint subsets, such as that appearing in chapter 5, could be used to directly estimate repeatability without bias in a real dataset.

10.9.3 Extension to consider multiple learners

The bolstering rule could be used to extend other statistical procedures that allow for the comparison of three or more learners. As described in section 10.4.2, care must be taken when selecting the statistic to be averaged over base CV experiment repetitions. I suggest the following possible directions:

- The L performances of L learners may be considered a vector. The set of performances observed on an item may be modelled as a random vector from a multivariate normal distribution whose mean is the true performances of those learners. In the single predictor case, one can use the sample variance and mean to construct a confidence region based on Hotelling's T^2 distribution. One could construct a bolstered test where the sample mean and covariance used to determine were smoothed over multiple experiment repetitions. The multivariate normal model can to some extent be justified by the applicability of central limit theorems in the single predictor case. The same model could be used to provide a test against the null hypothesis that all learners have equal performance.

- In classification problems, where performance measurements are binary, one may use Cochran's Q test to test against the null hypothesis in that all learners have equal performance. In a manner equivalent to that seen in the bolstered techniques based on binomial

models, the average performances on an item taken over multiple experiments could be used in the place of the binary measurement expected from a single experiment.

- The Friedman test is a non-parametric test commonly used to detect differences in the performances of multiple predictors under the single predictor model. It is generally applicable, though it does not strictly test against a null hypothesis of equal means. Rather, it tests against a null hypothesis that all rankings are equally likely. In a bolstered version of the test to compare multiple predictors, the sums of squares that form the numerator and denominator of its test statistic could be smoothed over multiple experiment repetitions.

One remaining challenge for all tests to detect differences in mean performance would be the construction of *post hoc* tests.

Summary

I have introduced the voting and bolstering conservative extension rules for the construction of heuristic statistical procedures. The new tests exploit the additional information provided by sequential KCV or SHOCV experiment repetitions on the same dataset without limiting assumptions about how much there is. This allows them to be used with low variance CV strategies such as RKCV without the need for potentially unreliable calibration.

I have analysed both rules from the perspective of increased repeatability (over baseline statistical procedures based on fixed predictor models in a single repetition of KCV or SHOCV). I have argued that repeatability is a good surrogate metric for statistical procedures, as high repeatability is associated with low type I error rates and high power at larger effect sizes. On the basis of my analysis, I have selected the bolstering rule for further validation.

In synthetic binary classification tasks, the bolstering rule increased the coverage of confidence intervals with only a minimal increase in expected width. The improvement in coverage was greatest in those cases where the problem of dependency caused the greatest reduction in coverage below the nominal values. The power of a pairwise comparison test was improved at large effect sizes, and its type I error rate was reduced.

AD classification experiments in the ADNI dataset show that inference based on a SHOCV, as appears sometimes in AD ANA, can have low replicability, which itself implies low power. Statistical procedures based on the bolstering rule can remove this limitation.

The bolstering rule could be used to allow inference using low variance CV strategies in AD ANA. This inference will be by design more reliable than the inference based on single KCV and SHOCV repetitions which is currently practised.

Part V

Conclusions

Chapter 11

Conclusions

In this thesis, I have offered a unique review of key technical validation problems facing automated neuroimaging assessment (ANA) research for Alzheimer's disease (AD) and made several novel contributions that may help to solve them. This chapter summarises the work of this thesis, discusses its implications, and points to future areas of extension.

11.1 Selection bias

The study of chapter 5 demonstrates that selection bias can plausibly account for a significant fraction of the apparent performance improvement associated with learner specification optimisation, even at realistic sample sizes. As expected under the simple Gaussian measurement model of chapter 4, variance in performance measurement and the number of learners considered for selection are the crucial determinants of selection bias. Even in a relatively low variance setting such as AD detection with 300 subjects, selection bias was responsible for more than 20% of apparent improvement. In the prediction of conversion from mild cognitive impairment (MCI), where samples are smaller and variance is intrinsically higher, bias accounted for more than two thirds of the observed improvement even at the maximum sample size of 160.

A key finding of the study is that, when performances are reported selectively, smaller sample sizes may actually be associated with higher performance estimates. Where one would expect larger samples to be association with training set sizes, and so with greater performance, the greater potential for bias associated with small samples can actually overwhelm this effect.

The observations of the study are consistent with two observations from the literature. The first of these is that the in-sample performance estimates reported in challenges are almost always more optimistic than the corresponding unbiased out of sample performance estimates, which shows that learner optimisation is occurring in those contexts. The second is the apparent negative association between sample size and reported performance in the AD classification literature.

While my study and findings are novel, my proposed solutions are not. Selection bias is an inevitable consequence of the search for the best learner specifications in the absence of unlimited data. It cannot be eliminated, but it can be reduced and its negative effects can be mitigated. Low variance strategies such as repeated K-fold cross validation (RKCV) with a high numbers of K-fold repetitions can reduce bias appreciably, and lower selections of the parameter K can be reduce variance without increasing computation cost. Maximising sample size before future learner selection is key, and performance estimates from previous small-sample studies should be interpreted with care. Challenges and other mechanisms of more transparent reporting can make the level of selection occurring more transparent, and thus allow interpreters to gauge when selection bias is more likely.

11.1.1 Future work

A replication study which implemented and validated various high performing methods from the literature could also be valuable in assessing selection bias. The one limitation of such a study is that the necessary overlap between the sample it used and the samples used in the studies to be replicated might cause random effects to be shared between them.

11.2 Cross validation strategies

As discussed in chapter 6, high precision and low bias are desirable properties of a cross validation (CV) strategy for use in AD ANA. High precision is particularly important when selecting learners. Variance can be reduced by using additional train-test experiments or by increasing efficiency. Equal use CV strategies are to be preferred for their greater efficiency. Stratification and similar strategies should be used to reduce bias and variance, particularly in the stratification problems that make up the bulk of those studied in AD ANA.

The extended K-fold cross validation strategy (EKCV) I have demonstrated is a useful generalisation of K-fold cross validation (KCV) to allow for a larger number of training set sizes while retaining the equal use criterion. It is particularly useful in experiments where this must be precisely controlled.

In general, smaller training set sizes (less than or equal to two thirds of the available items) are to be preferred in RKCV and EKCV for the lower variance they can provide using a fixed number of train-test experiments. Even where one wishes to assess the performance of a learner on a training set equal in size to the full sample, the lower variance associated with small training sets more than compensates for the increased bias. In learner selection, where bias is of less relative importance, the incentive for smaller training set sizes is even greater.

The balanced incomplete cross validation (BICV) strategy described in [135] has the po-

tential to allow for more efficient CV than the KCV and RKCV that are typically used. Unfortunately, BICV places severe restrictions on the training set sizes that can be used. It also cannot be used with stratification, which is particularly important in the classification problems most common in AD ANA. In chapter 7, I developed approximately balanced cross validation (ABCV) to overcome these limitations. ABCV is an approximation of BICV based on a greedy selection of training sets. Preliminary experiments show that while ABCV is indeed more efficient than RKCV, the gain in efficiency is so small as to be practically insignificant. For now, RKCV is still an essentially optimal strategy in problems where stratification is important.

11.2.1 Future work

The algorithm of ABCV could be improved by changing the algorithm used to find optimal block designs.

11.3 Statistical procedures for cross validation

The problem of dependency (described in chapter 8) means that conventional statistical procedures may fail when applied to make statements about learner performances based on the results of CV. In practice this means that type I error rates, corresponding to false positives in significance tests and non-containment of the true parameter in interval estimation, may be greater than their nominal values.

Concerns about the strict validity of statistical procedures may be the reason that statistical analysis is largely absent in the AD ANA literature. This is regrettable, as it makes it difficult to interpret the results of many studies and to assess how confident one should be in any estimation of performance. Some flawed analysis should be preferred over point estimation alone, and there is a motivation for new statistical procedures better suited for use in CV.

Various approaches to the problem of inference in CV have been proposed in the last two decades. As discussed in chapter 9, many of these share crucial flaws including undesirable calibration parameters that limit the plausible validity of a procedure to problems similar to those used in calibration, an incompatibility with low variance CV strategies, an incompatibility with binomial models, or a limited replicability and power. This motivates the development of new procedures.

In chapter 10, I have introduced the idea of a conservative extension rule for the construction of new statistical tests. These abandon the aim of producing exact p values and interval coverages, and instead produce procedures that are assuredly better than 'baseline' tests based on fixed predictor assumptions. This improvement is measured in terms of greater repeatability, which implies lower type I error rates and greater power at large effect sizes. This improve-

ment is achieved through the incorporation of an unknown amount of additional information provided by additional repetitions of a simple CV experiment. I have offered a brief theoretical analysis of two extension rules called voting and bolstering, and selected the bolstering rule for further validation.

The bolstering rule is based on a combination of mean test statistics across CV experiment repetitions. It should provide an increase in repeatability whenever the relevant test statistics are approximately normal, an assumption shared by the vast majority of statistical procedures. It should offer the greatest improvements when deviations from the fixed predictor model are high, which is the precisely when specialist statistical procedures are needed the most.

Validation in synthetic classification problems demonstrates that bolstering can increase the coverage of confidence intervals with minimal increases in their expected widths, with the greatest gains seen where coverage was most reduced below the nominal value by the problem of dependency. Bolstered pairwise tests for differences in performance between two learners had lower type I error rates and greater power at larger effect sizes. A small study on AD classification demonstrates that bolstering can be used to dramatically improve the replicability of the inference based on simple hold-out cross validation (SHOCV) that may be considered by some as a gold standard. The low replicability of the baseline SHOCV-based method points to low power without bolstering.

11.3.1 Future work

A more detailed study would compare the bolstered inference procedures to all comparable alternatives from the literature in a range of real and synthetic problems. This could include datasets from standard machine learning repositories in addition to neuroimaging data.

11.4 Centralised validation

Ultimately, the work of this thesis has led me to believe that the current research paradigm is not well suited to the goal of identifying learner specifications with higher performance. If the research community it to efficiently pursue this goal, there will need to be more standardised and centralised evaluation of proposed learner specifications, as occurs in challenges.

In particular, I suggest a grand challenge in which contestants submit learners, rather than predictions. The organisers of this challenge would use a low variance CV strategy to validate these on a large hidden dataset, and all performance results would be reported. This would allow all learner specifications to be compared under identical conditions, removing the differences that normally confound comparison between studies.

A grand challenge of the type described would report all performance results regardless of

whether or not they were impressive. The estimates would individually be unbiased this way, and the challenge organisers could use an appropriate confidence interval procedure to produce estimates of uncertainty for all submissions. Because all results would be reported, it would be easier to anticipate the level of selection bias associated with the highest performance results.

Through a simple interface, such a project could provide contestants access to a variety of processed images and features, thus lifting the burden of computation and implementation from researchers whose processing steps may not be unique. Because the imaging data themselves would never have to be distributed, the project could also use large samples without concern for data protection issues. In addition to the unbiased performance estimates for submitted learners, the project could also simultaneously implement the unbiased selection strategy described in [143]. This would provide an unbiased estimate of the best performance the research community can produce if they use CV to select the best available learner specification from the set of submitted alternatives.

Appendices

Appendix A

On the bias implications of initial transformations in cross validation

As discussed in section 2.5.1, there are many cases where one may wish to perform some data-driven transformation of the features before applying a standard learning algorithm. This transformation might be some kind of standardisation (e.g., scaling real valued features to ensure unit variance) or dimensionality reduction (e.g., projection to independent components). Even groupwise registration and atlas propagations (see section 2.4) may be considered initial transformations of this kind. In CV experiments where these types of transformations are used, there are the two following possible **experimental setups**.

1. One can learn a single transformation on the full dataset and apply it to all items before conducting any CV experiments.

2. One can learn a new transformation for each individual train-test experiment that appears in CV. This is learned using the training set alone, and then applied to both training and testing sets.

This appendix is concerned with the effect of initial transformations on the expected performance of learners in train-test experiments and explains how this can introduce bias into CV performance measures. Section A.1 describes how the use or non-use of an item in transformation learning can affect its distribution in the transformed feature space. The changes in distribution induced this way can confound learning algorithms. Section A.2 discusses the practical implications of this in CV experiments, and considers when each experimental setup should be used in practice.

A.1 Distribution shift

It should be noted that, in the transformed feature space, *the distribution of the items that are used to learn the transformation may differ from the distribution of items that are not*. I shall

call this effect **distribution shift**. To illustrate how it may arise, consider the following simple example. Let the item features X comprise a single real value (that is, $\mathbb{X} = \mathbb{R}$) where $\mathbb{E}[X] = \mu$ and $\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mu^2 = \sigma^2$. The transformation to be learned is a simple demeaning operation, where transformed features X' are defined as $X' = X - \bar{X}$, where $\bar{X}$ is simply the mean of the m observations in a training set. It is straightforward to show that $\mathbb{E}[X'^2] = \sigma^2 m/(m-1)$ for those items used to learn the transformation, while $\mathbb{E}[X'^2] = \sigma^2(m+1)/m$ for independent items not used in its construction. This demonstrates that the distribution of transformed item features differs between the two item types. Note also that while the transformed features X' of the m items used to learn the transformation remain exchangeable (that is, their joint distribution is invariant to permutations), they are no longer independent.

Where only the items of a training set are used to learn a transformation, the resulting distribution shift can make the relationships between the new features and the labels differ between the training and testing sets. This difference violates a key assumption behind many learning algorithms that are typically applied at this stage, and so it may be detrimental to performance. Using the full set of items for transformation construction eliminates this issue by treating all items symmetrically and so ensuring there is no shift in distribution between the training and testing sets. This effect, along with the potentially more useful transformation informed by more items, means that *the first experimental setup will typically produce higher performance estimates*.

The *detrimental effect of distribution shift to predictive performance may be particularly profound when the transformation learning uses the items' labels*. These types of transformations are often chosen deliberately to induce a particular relationship between the transformed features and the labels in the items used to learn the transformation. One example in ANA is projection to partial least squares components [55], which involves the projection of a d Euclidian feature space into a $d' < d$ dimensional linear subspace so as to maximise the degree to which real valued labels can be predicted by a linear model. In cases where d is high relative to the number of available items m, good linear fits will exist in a transformation learning set *even if none truly exist in the items' distribution*. In this case, though the transformation will always be able to ensure a good linear relationship between the transformed features and labels for the items of the transformation learning set, this relationship will not persist when the transformation is applied to independent items. This will introduce a substantial drop in expected performance when moving from the first to the second experimental setup.

A.1.1 When does distribution shift matter?

While distribution shift is likely to have some effect for nearly all transformations, there are limits where it becomes less important. Many types of transformation will converge to some expected value determined by the distribution of items in the original feature space as $m \to \infty$. As this happens, the relationship between the transformation and the items the transformation learning set will weaken, and the difference between their distribution in the new feature space and that of independent items will vanish. An example of this is found in the simple demeaning operation considered in this appendix, where the difference in $\mathbb{E}[X'^2]$ between item sets is of the order of m^{-2}. Other transformations may converge too slowly to offer any guarantee that distribution shift can be ignored. Intuitively, convergence is likely to be slower when the transformation is specified by a higher number of parameters to be estimated.

One way to reduce the effect of distribution shift in CV would be to weaken the association between the transformation and a particular validation dataset set by learning the transformation with additional items that can not be used in validation. An example of this in ANA might be including many additional subjects when building a representative target image in groupwise registration, even though these may not be used for training and evaluating learners.

Naturally, the effect of distribution shift on the expected performance in a given train-test experiment will also be dependent on the learners under study. Different types of distribution shift will interact differently with different types of learning algorithm.

A.2 Practical implications in cross validation

In ANA and much other machine learning research, it is not typically expected that an initial transformation will be relearned using new unlabelled items as they arrive in the imagined future application of the learners under study. In this framework, it is the second experimental setup that is more realistic. *To use the first experimental setup for CV will create an unrealistic situation where the distribution shift that will be present in the future application is removed. This will risk giving the CV performance estimates an optimistic bias, particularly when the labels are required to learn transformations.* If one instead uses the second setup, then the transformation learning process can be considered part of the learners under study (and the application as part of their predictor, as discussed in section 3.1.1). Some pessimistic bias may be introduced where the setup forces one to reduce the number of items available for training further below the number anticipated in the future application. As discussed in section 4.2.2, pessimistic biases should be of less concern to ANA researchers than optimistic ones. For this reason, researchers should favour the second setup in the general case.

While the second setup should be generally preferred, there will be cases where the first may be permissible. Researchers must inevitably weigh the risk of optimistic bias against the limitations imposed by increased computational cost. If transformation learning does not use the item labels and is expected to be relatively stable, then the distribution shift may have a negligible effect on the estimated performance of learners, and the potential for optimistic bias is smaller. One need not, for instance, discount the results of every study that performed a demeaning and rescaling transformation before performing CV.

A.2.1 An exception to the rule

In cases where the labels are not used in transformation learning, one can imagine a hypothetical future application where unlabelled items *are* used to relearn a transformation each time they arrive. In this case, it is the first experimental setup that is more realistic, as no distribution shift is to be expected in future. This learning problem is not well described by the supervised learning formalism used in this thesis, and it has more in common with semi-supervised learning [202]. Note that, in a train-test experiment where this type of learning is applied, the expected performance will not only depend on the number of training items, but also on the number of testing items.

Summary

The distribution of an item's features in a transformed feature space may differ depending on whether it was used to construct the transformation. In CV experiments and real applications where only the training data are used to learn a transformation, this distribution shift can reduce predictive performance, particularly when the item labels are used to learn a transformation. If, in cross validation, items outside the training set are used to learn an initial transformation, then this can unrealistically lift a barrier to prediction that will be present in the real application. Although limiting the number of items used for transformation learning can limit performative accuracy, researchers must relearn their transformations separately on each training set that appears in a CV experiment to avoid this source of optimistic bias. For this reason, the second experimental setup should be preferred in the general case.

While the optimistic bias of distribution shift is always a potential problem, there are cases where it may not be important enough to justify the increased computational cost of the second experimental setup. In cases where transformation learning is sufficiently stable, the effect of distribution shift on the performance of learning algorithms will become small, and it will be permissible to use the first setup. Where applicable, using additional comparable items from outside the CV dataset for transformation learning may provide a way to further diminish dis-

tribution shift in these cases. Crucially, it will not generally be permissible to use the first setup with transformation learning processes that make use of the item labels, as this can introduce distribution shifts that interact strongly with learning algorithms.

Appendix B

Proof for decreased variance under stratification

Let μ_s denote the expected value of a performance estimate for a prediction on an item from subpopulation s in a testing experiment for a fixed predictor. Let β_s denote the expected second moment of such an estimate, and σ_s^2 denote the variance. Let f_s denote the fraction of the full population's probability mass associated with subpopulation s. Let S denote the total number of disjoint subpopulations that make up the full population. By definition, $\sum_{s=1}^{S} f_s = 1$. According to the standard formula for the variance of a random variable, $\sigma_s^2 = \beta_s - \mu_s^2$. The first and second moments of a performance estimate on an item drawn at random from the full population are

$$\mu = \sum_{s=1}^{S} f_s \mu_s \text{ and} \tag{B.1}$$

$$\beta = \sum_{s=1}^{S} f_s \beta_s \tag{B.2}$$

respectively. The variance σ^2 of the full population is given

$$\begin{aligned} \sigma^2 &= \beta - \mu^2 \\ &= \sum_{s=1}^{S} f_s \beta_s - \Big(\sum_{s=1}^{S} f_s \mu_s \Big)^2 . \end{aligned} \tag{B.3}$$

The variance v of the mean performance estimate in a non-stratified testing set of n i.i.d. items from the full population will therefore be

$$\begin{aligned} v &= \frac{\sigma^2}{n} \\ &= \frac{1}{n} \sum_{s=1}^{S} f_s \beta_s - \frac{1}{n} \Big(\sum_{s=1}^{S} f_s \mu_s \Big)^2 . \end{aligned} \tag{B.4}$$

This may be compared with a stratified testing set of $n = \sum_{s=1}^{S} n_s$ items, where $n_s = nf_s$ denotes the number of items from subpopulation s. In this case, the variance v' of the mean performance observed on the stratified testing set will be

$$\begin{aligned} v' &= \frac{1}{n^2} \sum_{s=1}^{S} n_s \sigma_s^2 \\ &= \frac{1}{n} \sum_{s=1}^{S} f_s \sigma_s^2 \\ &= \frac{1}{n} \sum_{s=1}^{S} f_s \beta_s - \frac{1}{n} \sum_{s=1}^{S} f_s \mu_s^2 \,. \end{aligned} \tag{B.5}$$

The difference between these two is proportional to

$$\begin{aligned} v - v' &= \frac{1}{n} \sum_{s=1}^{S} f_s \mu_s^2 - \frac{1}{n} \Big(\sum_{s=1}^{S} f_s \mu_s \Big)^2 \\ &= \frac{1}{n} \left[\sum_{s=1}^{S} f_s \mu_s^2 - 2 \Big(\sum_{s=1}^{S} f_s \mu_s \Big)^2 + \Big(\sum_{s=1}^{S} f_s \mu_s \Big)^2 \right] \\ &= \frac{1}{n} \sum_{s=1}^{S} f_s \Big[\mu_s^2 - 2\mu_s \Big(\sum_{s'=1}^{S} f_{s'} \mu_{s'} \Big) + \Big(\sum_{s'=1}^{S} f_{s'} \mu_{s'} \Big)^2 \Big] \\ &= \frac{1}{n} \sum_{s=1}^{S} f_s \Big[\mu_s - \Big(\sum_{s'=1}^{S} f_{s'} \mu_{s'} \Big) \Big]^2 . \end{aligned} \tag{B.6}$$

The summands in the final line of equation (B.6) are clearly non-negative. *This means that v', the variance of the performance estimate in a stratified testing set with representative subpopulation proportions, is always less than or equal to v, the variance in a non-stratified testing set.* Equality between v and v' holds if and only if $\mu_s = \mu$ for all s.

Appendix C

Proof of increased repeatability under majority vote

Where a variable X has a symmetric unimodal distribution on the interval $[0,1]$, this distribution may be divided into regions as in figure C.1. The lengths of these regions are denoted x_a and x_b, and their mean probability densities are denoted a and b. When $\mathbb{E}[X] > 1/2$,

$$\mathbb{E}[X] = 1/2 + x_b \text{ and} \tag{C.1}$$

$$P(X > 1/2) = 1/2 + b\dot{x}_b\,. \tag{C.2}$$

From the symmetry of the distribution,

$$b\dot{x}_b + a\dot{x}_a = 1/2 \tag{C.3}$$

From the unimodal property of the distribution $b > a$. Therefore,

$$b\dot{(}x_b + x_a) = 1/2\,. \tag{C.4}$$

Because the distribution of X is bounded between 0 and 1, $a + b < 1/2$. Therefore

$$b > 1\,. \tag{C.5}$$

Applying this result to equations (C.1) and (C.2) produces the result $P(X > 1/2) > \mathbb{E}[X]$.

By symmetry, $P(X > 1/2) < \mathbb{E}[X]$ when $\mathbb{E}[X] < 1/2$.

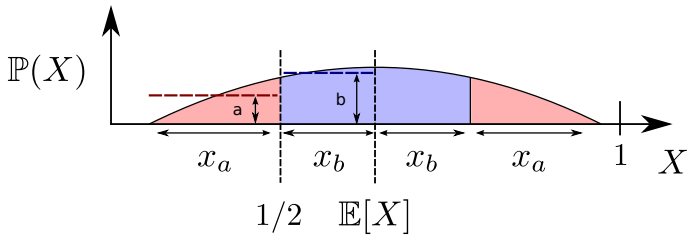

Figure C.1: A symmetric and univariate distribution of the latent test parameter X. The distribution may be described by two identical region pairs with mean probability densities a and b, and lengths x_a and x_b.

Appendix D

Proof of increased replicability under majority vote

Let Y denote the number of positive outcomes in a sequence of E statistical tests conducted on CV experiments associated with sequential random partitions on a sample with a latent rate parameter W. In the voting combination meta-heuristic, a final decision is reached by majority vote amongst their respect outcomes. Because Y has a binomial distribution conditional on W, its cumulative distribution may be written

$$P(Y \leq e) = B_{1-W}(E-e, e+1) \tag{D.1}$$

where the $B_x(a,b)$ is the regularised incomplete beta function defined

$$B_x(a,b) := \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt\,, \tag{D.2}$$

for $x \in [0,1]$ and $B(a,b)$ is the 'complete' beta function defined

$$B(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1} dt\,, \tag{D.3}$$

Let E be an odd number to avoid ties in voting. From equation (D.1), the chance of a negative outcome in a voting test using E repetitions is given

$$P_E^- = B_{1-W}\Big(\frac{E+1}{2}, \frac{E+1}{2}\Big)\,, \tag{D.4}$$

While that for $E+2$ repetitions is

$$P_{E+2}^- = B_{1-W}\Big(\frac{E+1}{2}+1, \frac{E+1}{2}+1\Big)\,. \tag{D.5}$$

Using the substitutions $a = (E+1)/2$ and $x' = 1-W$, one may write $P_E^- = B_{x'}(a,a)$ and $P_{E+2}^- = B_{x'}(a+1,a+1)$. Applying lemma 1, one can demonstrate that

$$\begin{aligned} P_{E+2}^- &> P_E^- \text{ for } W < \frac{1}{2}, \\ &= P_E^- \text{ for } W = \frac{1}{2}, \\ &< P_E^- \text{ for } W > \frac{1}{2}. \end{aligned} \tag{D.6}$$

This means that increasing the number of repetitions in a voting test on a given dataset increases the probability of the most likely decision given that dataset (for $W \neq 1/2$). Necessarily, all voting-based tests using odd E are more likely to produce the most likely result than the baseline test, which effectively uses $E = 1$. It also follows that, where the chance of getting the same result in any two sequential tests is given

$$\rho_E = (P_E^-)^2 + (1-P_E^-)^2, \tag{D.7}$$

it can be guaranteed that

$$\rho_{E+2} \geq \rho_E, \tag{D.8}$$

with equality holding only in the case of $W = 1/2$. Because replicability with E repetitions may be defined

$$\text{replicability}_E = \mathbb{E}_{\rho_E}[W], \tag{D.9}$$

it is straightforward to show that

$$\text{replicability}_{E+2} > \text{replicability}_E \tag{D.10}$$

whenever $W \neq 1/2$ with non-zero probability.

Lemma 1. *For $a \geq 0$, the sign of $x - 1/2$ determines the relative sizes of $B_x(a,a)$ and $B_x(a+1,a+1)$ in the following way:*

$$\begin{aligned} B_x(a+1,a+1) &> B_x(a,a) \textit{ for } x > \frac{1}{2}, \\ &= B_x(a,a) \textit{ for } x = \frac{1}{2}, \\ &< B_x(a,a) \textit{ for } x < \frac{1}{2}. \end{aligned} \tag{D.11}$$

Proof. The following are identities from Abramowitz and Stegun [203]:

$$B_x(a+1,b) = B_x(a,b) - \frac{x^a(1-x)^b}{aB(a,b)}, \tag{D.12}$$

$$B_x(a,b+1) = B_x(a,b) + \frac{x^a(1-x)^b}{bB(a,b)}, \tag{D.13}$$

$$aB(a,b) = (a+b)B(a+1,b). \tag{D.14}$$

These can be combined to show that

$$\begin{aligned}
B_x(a+1,b+1) &= B_x(a+1,b) + \frac{x^{a+1}(1-x)^b}{bB(a+1,b)} && \text{from (D.12),} \\
&= B_x(a,b) + \frac{x^{a+1}(1-x)^b}{bB(a+1,b)} - \frac{x^a(1-x)^b}{aB(a,b)} && \text{from (D.13),} \\
&= B_x(a,b) + x^a(1-x)^b\left(\frac{x}{bB(a+1,b)} - \frac{1}{aB(a,b)}\right) \\
&= B_x(a,b) + x^a(1-x)^b\left(\frac{x}{bB(a+1,b)} - \frac{1}{(a+b)B(a+1,b)}\right) && \text{from (D.14).}
\end{aligned} \tag{D.15}$$

In the case $a = b$, this may be expressed as

$$B_x(a+1,a+1) = B_x(a,a) + \frac{x^a(1-x)^a}{aB(a+1,a)}\left(x - \frac{1}{2}\right) \tag{D.16}$$

Where $a \geq 0$, and $B(a+1,a) \geq 0$, this means that the sign of the rightmost summand is determined solely by the sign of $x - 1/2$.

□

Appendix E

Reduction in absolute central moments under averaging

This chapter presents a proof I encountered in [204]. It demonstrates that the average of a series of exchangeable real valued random variables has absolute central moments no greater than those of any one of those variables, regardless of the variables' dependency structure.

Let $\langle X_i \rangle_{1 \leq i \leq E} \in \mathbb{E}^E$ denote a sequence of random variables that share a marginal mean μ. The average of these is

$$\bar{X} = \frac{1}{E} \sum_{e=1}^{E} X_e \,. \tag{E.1}$$

An absolute central moment of a random variable X is given $\mathbb{E}[\,|X - \mu|^z]$, where $\mu = \mathbb{E}[X]$. Where the X may be viewed as an estimator for $\mathbb{E}[X]$, this definition includes a broad family of error measures. Jensen's inequality for a convex function $f : \mathbb{E} \to \mathbb{E}$ may be written

$$f\Big(\frac{x_1 + x_2 + \cdots + x_E}{E}\Big) \leq \frac{f(x_1) + f(x_2) + \cdots + f(x_E)}{E} \,. \tag{E.2}$$

Because $|X - \mu|^z$ is convex for all $z > 1$, one can apply the inequality to produce

$$\begin{aligned} |\bar{X} - \mu|^z &= \left| \frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_E - \mu)}{E} \right|^z \\ &\leq \frac{|X_1 - \mu|^z + |X_2 - \mu|^z + \cdots + |X_E - \mu|^z}{E} \,. \end{aligned} \tag{E.3}$$

Taking the expectation of both sides, and using the fact that $\mathbb{E}[\,|Q_e - Q|^z]$ is the same for all e, produces the following:

$$\mathbb{E}\big[\,|\bar{X} - \mu|^z\big] \leq \mathbb{E}\big[\,|X_e - \mu|^z\big] \,, \text{ for all } z > 1. \tag{E.4}$$

This demonstrates that $\bar{X}$ will have lower absolute central moments than any individual X_e,

including a lower variance.

Appendix F

Replication of earlier study

Figure F.1 presents a replication of the study described in chapter 10.7 where RHOCV was used in place of RKCV. As before, stratification was used to select training sets comprising half the available items. Results are qualitatively similar to those seen with RKCV, but the improvement seen with bolstering is more dramatic.

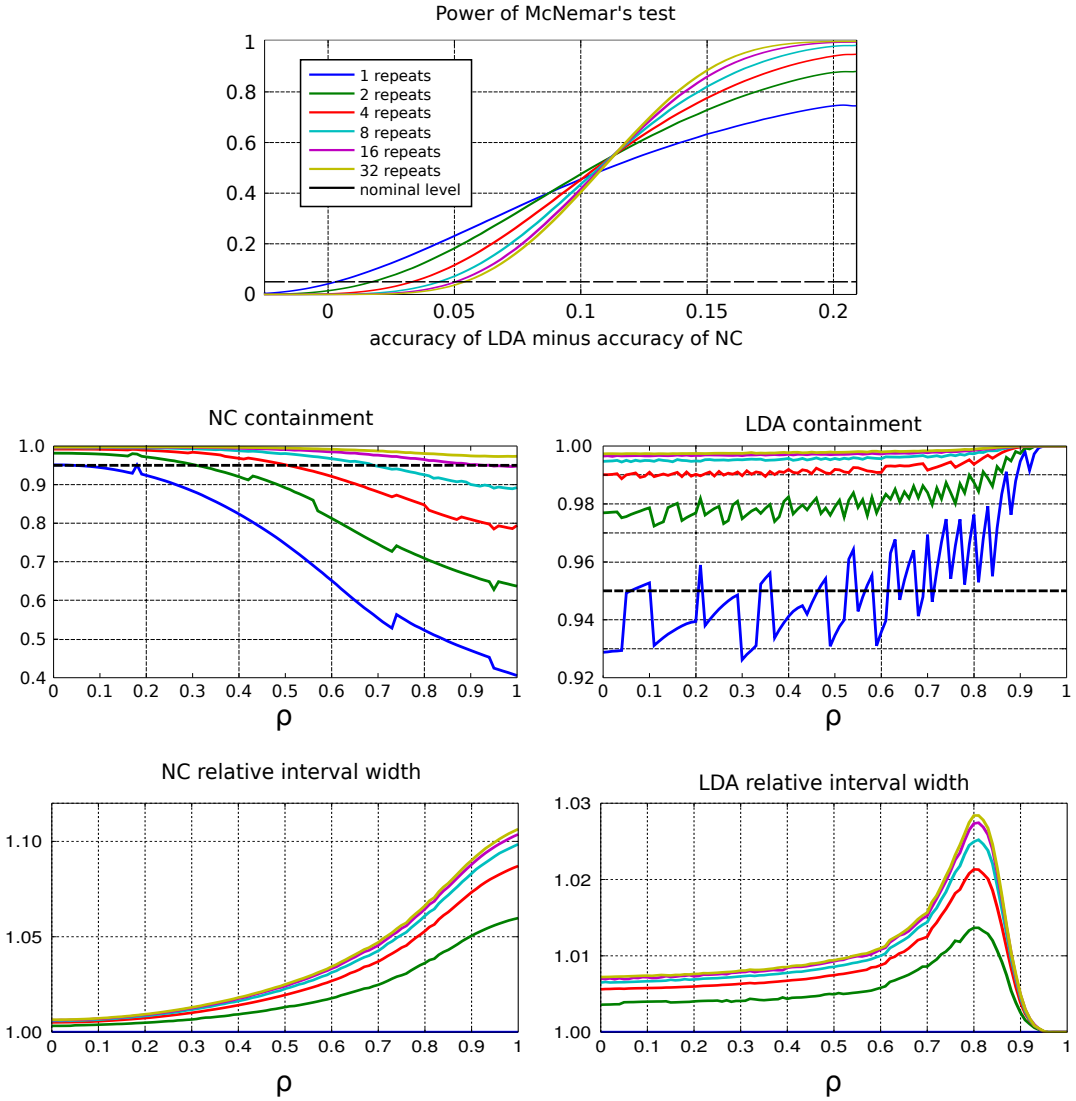

Figure F.1: Power of bolstered McNemar's test against the null hypothesis that the LDA learner has a performance no better than that of the NC learner (above). Interval coverage rates for NC and LDA learner performances (vertical centre). Expected width of confidence intervals relative to those of the single SHOCV repetition procedure (below).

Bibliography

[1] Martin Prince, Renata Bryce, Emiliano Albanese, Anders Wimo, Wagner Ribeiro, and Cleusa P. Ferri. The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1):63 – 75.e2, 2013.

[2] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack Jr., Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. The diagnosis of dementia due to Alzheimers disease: Recommendations from the national institute on aging-Alzheimers association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263 – 269, 2011.

[3] Alistair Burns and Steve Iliffe. Dementia. *BMJ*, 338, 2009.

[4] Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H. Michael Arrighi. Forecasting the global burden of Alzheimers disease. *Alzheimer's & Dementia*, 3(3):186 – 191, 2007.

[5] David L Weimer and Mark A Sager. Early identification and treatment of Alzheimer's disease: social and fiscal outcomes. *Alzheimer's & Dementia*, 5(3):215–226, 2009.

[6] Anders Wimo, Linus Jnsson, John Bond, Martin Prince, and Bengt Winblad. The worldwide economic impact of dementia 2010. *Alzheimer's & Dementia*, 9(1):1 – 11.e3, 2013.

[7] Alex F Mendelson. A list of publications describing new supervised learning pipelines to predict clinical variables from neuroimaging data in Alzheimer's disease. 2016.

[8] Bradley T Hyman, John H Growdon, Mark W Albers, Randy L Buckner, Jasmeer Chhatwal, Maria Teresa Gomez-Isla, Christian Haass, Eloise Hudry, Clifford R Jack, Keith A Johnson, et al. Massachusetts Alzheimer's disease research center: Progress

and challenges. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(10):1241–1245, 2015.

[9] Kaj Blennow and Henrik Zetterberg. Cerebrospinal fluid biomarkers for Alzheimer's disease. *Journal of Alzheimer's Disease*, 18(2):413–417, 2009.

[10] Stephen Salloway, Reisa Sperling, Nick C. Fox, Kaj Blennow, William Klunk, Murray Raskind, Marwan Sabbagh, Lawrence S. Honig, Anton P. Porsteinsson, Steven Ferris, Marcel Reichert, Nzeera Ketter, Bijan Nejadnik, Volkmar Guenzler, Maja Miloslavsky, Daniel Wang, Yuan Lu, Julia Lull, Iulia Cristina Tudor, Enchi Liu, Michael Grundman, Eric Yuen, Ronald Black, and H. Robert Brashear. Two phase 3 trials of Bapineuzumab in mild-to-moderate Alzheimer's disease. *New England Journal of Medicine*, 370(4):322–333, 2014. PMID: 24450891.

[11] Andrew P Prayle, Matthew N Hurley, and Alan R Smyth. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *Bmj*, 344:d7373, 2012.

[12] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

[13] Eric Budish, Benjamin N Roin, and Heidi Williams. Do firms underinvest in long-term research? Evidence from cancer clinical trials. Technical report, National Bureau of Economic Research, 2013.

[14] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[15] Clifford R Jack, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013.

[16] Eric Karran, Marc Mercken, and Bart De Strooper. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nature reviews Drug discovery*, 10(9):698–712, 2011.

[17] SM Landau, BA Thomas, L Thurfjell, M Schmidt, R Margolin, M Mintun, M Pontecorvo, SL Baker, WJ Jagust, Alzheimers Disease Neuroimaging Initiative, et al. Amyloid PET imaging in Alzheimers disease: a comparison of three radiotracers. *European journal of nuclear medicine and molecular imaging*, 41(7):1398–1407, 2014.

[18] PH Scheltens, D Leys, F Barkhof, D Huglo, HC Weinstein, P Vermersch, M Kuiper, M Steinling, E Ch Wolters, and J Valk. Atrophy of medial temporal lobes on MRI in probable Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(10):967–972, 1992.

[19] Lorena Bresciani, Roberta Rossi, Cristina Testa, Cristina Geroldi, Samantha Galluzzi, Mikko P Laakso, Alberto Beltramello, Hilkka Soininen, and Giovanni B Frisoni. Visual assessment of medial temporal atrophy on MR films in Alzheimers disease: comparison with volumetry. *Aging clinical and experimental research*, 17(1):8–13, 2005.

[20] Claire Boutet, Marie Chupin, Olivier Colliot, Marie Sarazin, Gurkan Mutlu, Aurélie Drier, Audrey Pellot, Didier Dormont, Stéphane Lehéricy, Alzheimers Disease Neuroimaging Initiative, et al. Is radiological evaluation as good as computer-based volumetry to assess hippocampal atrophy in Alzheimers disease? *Neuroradiology*, 54(12):1321–1330, 2012.

[21] Steven Ng, Victor L Villemagne, Sam Berlangieri, Sze-Ting Lee, Martin Cherk, Sylvia J Gong, Uwe Ackermann, Tim Saunder, Henri Tochon-Danguy, Gareth Jones, et al. Visual assessment versus quantitative assessment of 11C-PIB PET and 18F-FDG PET for detection of Alzheimer's disease. *Journal of Nuclear Medicine*, 48(4):547–552, 2007.

[22] Mert R Sabuncu, Ender Konukoglu, Alzheimers Disease Neuroimaging Initiative, et al. Clinical prediction from structural brain MRI scans: a large-scale empirical study. *Neuroinformatics*, 13(1):31–46, 2015.

[23] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Mller. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387 – 399, 2011. Multivariate Decoding and Brain Reading.

[24] Farshad Falahati, Eric Westman, and Andrew Simmons. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's Disease*, 41(3):685–708, 2014.

[25] Esther E Bron, Marion Smits, Wiesje M Van Der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M Papma, Rebecca ME Steketee, Carolina Méndez Orellana, Rozanna Meijboom, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, 111:562–579, 2015.

[26] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4):198–211, 2007.

[27] Stefan Klöppel, Ahmed Abdulkadir, Clifford R. Jack Jr., Nikolaos Koutsouleris, Janaina Mouro-Miranda, and Prashanthi Vemuri. Diagnostic neuroimaging across diseases. *NeuroImage*, 61(2):457 – 463, 2012.

[28] Jonathan Young, Marc Modat, Manuel J Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.

[29] Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137–165, 2017.

[30] Joseph Kambeitz, Lana Kambeitz-Ilankovic, Stefan Leucht, Stephen Wood, Christos Davatzikos, Berend Malchow, Peter Falkai, and Nikolaos Koutsouleris. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, 40(7):1742–1751, 2015.

[31] Ling-Li Zeng, Hui Shen, Li Liu, Lubin Wang, Baojuan Li, Peng Fang, Zongtan Zhou, Yaming Li, and Dewen Hu. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135(5):1498–1507, 2012.

[32] Graziella Orr, William Pettersson-Yeo, Andre F. Marquand, Giuseppe Sartori, and Andrea Mechelli. Using support vector machine to identify imaging biomarkers of neuro-

logical and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews*, 36(4):1140 – 1152, 2012.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[34] Stefan Klöppel, Cynthia M Stonnington, Carlton Chu, Bogdan Draganski, Rachael I Scahill, Jonathan D Rohrer, Nick C Fox, Clifford R Jack, John Ashburner, and Richard SJ Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[35] AF Mendelson, MA Zuluaga, L Thurfjell, BF Hutton, and S Ourselin. The empirical variance estimator for computer aided diagnosis: lessons for algorithm validation. In *Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 17, Pt 2, pages 236–243, 2014.

[36] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[37] Claude Nadeau and Yoshua Bengio. Inference for the Generalization Error. *Machine Learning*, 52(3), 2003.

[38] Yoshua Bengio and Yves Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.*, 5, December 2004.

[39] Michela Antonelli, Pietro Ducange, and Francesco Marcelloni. Genetic training instance selection in multiobjective evolutionary fuzzy systems: A coevolutionary approach. *Fuzzy Systems, IEEE Transactions on*, 20(2):276–290, 2012.

[40] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[41] John PA Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[42] David J Hand et al. Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14, 2006.

[43] Ahmed Abdulkadir, Bénédicte Mortamet, Prashanthi Vemuri, Clifford R Jack, Gunnar Krueger, Stefan Klöppel, Alzheimer's Disease Neuroimaging Initiative, et al. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. *NeuroImage*, 58(3):785–792, 2011.

[44] Simon F. Eskildsen, Pierrick Coup, Daniel Garca-Lorenzo, Vladimir Fonov, Jens C. Pruessner, and D. Louis Collins. Prediction of alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, 65:511 – 521, 2013.

[45] Katherine R Gray, Paul Aljabar, Rolf A Heckemann, Alexander Hammers, and Daniel Rueckert. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65, 2013.

[46] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen. Multi-modal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3), 2011.

[47] Xiaofeng Zhu, Heung-Il Suk, Li Wang, Seong-Whan Lee, and Dinggang Shen. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Medical Image Analysis*, pages –, 2015.

[48] Hervé Abdi, Lynne J Williams, Derek Beaton, Mette T Posamentier, Thomas S Harris, Anjali Krishnan, and Michael D Devous Sr. Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, frontotemporal dementia, and elderly controls: a multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer's Disease*, 31(s3), 2012.

[49] Alexander V Lebedev, E Westman, MK Beyer, MG Kramberger, C Aguilar, Z Pirtosek, and D Aarsland. Multivariate classification of patients with Alzheimers and dementia with Lewy bodies using high-dimensional cortical thickness measurements: an MRI surface-based morphometric study. *Journal of neurology*, 260(4):1104–1115, 2013.

[50] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895 – 907, 2012.

[51] Guan Yu, Yufeng Liu, and Dinggang Shen. Graph-guided joint prediction of class label and clinical scores for the Alzheimer's disease. *Brain Structure and Function*, pages 1–15, 2015.

[52] Kim-Han Thung, Pew-Thian Yap, Ehsan Adeli-M, and Dinggang Shen. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, chapter Joint Diagnosis and Conversion Time Prediction ofProgressive Mild Cognitive Impairment (pMCI) Using Low-Rank Subspace Clustering and Matrix Completion, pages 527–534. Springer International Publishing, Cham, 2015.

[53] Mert R. Sabuncu. *Machine Learning in Medical Imaging: 4th International Workshop, MLMI 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013. Proceedings*, chapter A Bayesian Algorithm for Image-Based Time-to-Event Prediction, pages 74–81. Springer International Publishing, Cham, 2013.

[54] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I*, chapter A Hybrid of Deep Network and Hidden Markov Model for MCI Identification with Resting-State fMRI, pages 573–580. Springer International Publishing, Cham, 2015.

[55] Anders H Andersen, William S Rayens, Yushu Liu, and Charles D Smith. Partial least squares for discrimination in fMRI data. *Magnetic resonance imaging*, 30(3):446–452, 2012.

[56] Wook Lee, Byungkyu Park, and Kyungsook Han. Classification of diffusion tensor images for the early detection of Alzheimer's disease. *Computers in Biology and Medicine*, 43(10):1313 – 1320, 2013.

[57] Katherine R. Gray, Robin Wolz, Rolf A. Heckemann, Paul Aljabar, Alexander Hammers, and Daniel Rueckert. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage*, 60(1):221 – 229, 2012.

[58] Carlos Cabral, Pedro M. Morgado, Durval Campos Costa, and Margarida Silveira. Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Computers in Biology and Medicine*, 58:101 – 109, 2015.

[59] R. Chaves, J. Ramirez, J. M. Gorriz, I. A. Illn, and D. Salas-Gonzalez. FDG and PIB biomarker PET analysis for the Alzheimer's disease detection using association rules. In *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*, pages 2576–2579, Oct 2012.

[60] J. Ramirez, J.M. Gorriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, M. Lpez, I. lvarez, and P. Padilla. Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. *Neuroscience Letters*, 472(2):99 – 103, 2010.

[61] Marie-Odile Habert, Jean-Franois Horn, Marie Sarazin, Jean-Albert Lotterie, Michle Puel, Fannie Onen, Michel Zanca, Florence Portet, Jacques Touchon, Marc Verny, Florence Mahieux, Alain Giron, Bernard Fertil, and Bruno Dubois. Brain perfusion SPECT with an automated quantitative tool can identify prodromal Alzheimer's disease among patients with mild cognitive impairment. *Neurobiology of Aging*, 32(1):15 – 23, 2011.

[62] Yang Li, Yaping Wang, Guorong Wu, Feng Shi, Luping Zhou, Weili Lin, and Dinggang Shen. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiology of Aging*, 33(2):427.e15 – 427.e30, 2012.

[63] Kim-Han Thung, Chong-Yaw Wee, Pew-Thian Yap, and Dinggang Shen. Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. *Brain Structure and Function*, pages 1–17, 2015.

[64] Junghoe Kim and Jong-Hwan Lee. Integration of structural and functional magnetic resonance imaging improves mild cognitive impairment detection. *Magnetic resonance imaging*, 31(5):718–732, 2013.

[65] Lele Xu, Xia Wu, Kewei Chen, and Li Yao. Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment. *Computer Methods and Programs in Biomedicine*, 122(2):182 – 190, 2015.

[66] Nikhil Singh, Angela Y. Wang, Preethi Sankaranarayanan, P. Thomas Fletcher, and Sarang Joshi. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part I*, chapter Genetic, Structural and Functional Imaging Biomarkers for Early

Detection of Conversion from MCI to AD, pages 132–140. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[67] Igor O. Korolev, Laura L. Symonds, Andrea C. Bozoki, and Alzheimer's Disease Neuroimaging Initiative. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS ONE*, 11(2):1–25, 02 2016.

[68] Francesca Mangialasche, E Westman, Miia Kivipelto, J-S Muehlboeck, R Cecchetti, M Baglioni, R Tarducci, G Gobbi, P Floridi, H Soininen, et al. Classification and prediction of clinical diagnosis of Alzheimer's disease based on MRI and plasma measures of α-/γ-tocotrienols and γ-tocopherol. *Journal of internal medicine*, 273(6):602–621, 2013.

[69] Yong Fan. *Multimodal Brain Image Analysis: First International Workshop, MBIA 2011, Held in Conjunction with MICCAI 2011, Toronto, Canada, September 18, 2011. Proceedings*, chapter Ordinal Ranking for Detecting Mild Cognitive Impairment and Alzheimer's Disease Based on Multimodal Neuroimages and CSF Biomarkers, pages 44–51. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[70] Fermn Segovia, Christine Bastin, Eric Salmon, Juan Manuel Gorriz, Javier Ramirez, and Christophe Phillips. Combining PET images and neuropsychological test data for automatic diagnosis of Alzheimer's disease. *PLoS ONE*, 9(2):1–8, 02 2014.

[71] Andrew Simmons, Eric Westman, Sebastian Muehlboeck, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Lars-Olof Wahlund, Hilkka Soininen, Simon Lovestone, et al. The AddNeuroMed framework for multi-centre MRI assessment of Alzheimer's disease: experience from the first 24 months. *International journal of geriatric psychiatry*, 26(1):75–82, 2011.

[72] Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T. Lautenschlager, Nat Lenzo, Ralph N. Martins, Paul Maruff, Colin Masters, Andrew Milner, Kerryn Pike, Christopher Rowe, Greg Savage, Cassandra Szoeke, Kevin Taddei, Victor Villemagne, Michael Woodward, and David Ames. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21:672–687, 8 2009.

[73] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[74] Francisco P.M. Oliveira and Joo Manuel R.S. Tavares. Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014. PMID: 22435355.

[75] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3), 2010.

[76] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the demons algorithm: 3D non-rigid registration by gradient descent. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI99*, pages 597–605. Springer, 1999.

[77] Torsten Rohlfing, Robert Brandt, Calvin R Maurer Jr, and Randolf Menzel. Bee brains, B-splines and computational democracy: Generating an average shape atlas. In *Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on*, pages 187–194. IEEE, 2001.

[78] Kenichi Ota, Naoya Oishi, Kengo Ito, Hidenao Fukuyama, SEAD-J Study Group, Alzheimer's Disease Neuroimaging Initiative, et al. Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease. *Journal of neuroscience methods*, 256:168–183, 2015.

[79] M. Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C. Fox, and Sebastien Ourselin. STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical Image Analysis*, 17(6), 2013.

[80] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, and Daniel Rueckert. LEAP: Learning embeddings for atlas propagation. *NeuroImage*, 49(2):1316 – 1325, 2010.

[81] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, Sept 2015.

[82] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of MR images of the brain. *Medical Imaging, IEEE Transactions on*, 18(10):897–908, 1999.

[83] M. Jorge Cardoso, Matthew J. Clarkson, Gerard R. Ridgway, Marc Modat, Nick C. Fox, and Sebastien Ourselin. LoAd: A locally adaptive cortical segmentation algorithm. *NeuroImage*, 56(3):1386 – 1397, 2011.

[84] John Ashburner and Karl J Friston. Voxel-based morphometrythe methods. *NeuroImage*, 11(6):805–821, 2000.

[85] Bruce Fischl and Anders M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97(20):11050–11055, 2000.

[86] Chloe Hutton, Enrico De Vita, John Ashburner, Ralf Deichmann, and Robert Turner. Voxel-based cortical thickness measurements in MRI. *NeuroImage*, 40(4):1701–1710, 2008.

[87] Chloe Hutton, Bogdan Draganski, John Ashburner, and Nikolaus Weiskopf. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, 48(2):371–380, 2009.

[88] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011.

[89] Mert R Sabuncu and Koen Van Leemput. The relevance voxel machine (RVOXM): A self-tuning bayesian model for informative image-based prediction. *Medical Imaging, IEEE Transactions on*, 31(12):2290–2306, 2012.

[90] Eric Westman, J-Sebastian Muehlboeck, and Andrew Simmons. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage*, 62(1):229 – 238, 2012.

[91] Emilie Gerardin, Gal Chtelat, Marie Chupin, Rmi Cuingnet, Batrice Desgranges, Ho-Sung Kim, Marc Niethammer, Bruno Dubois, Stphane Lehricy, Line Garnero, Francis

Eustache, and Olivier Colliot. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476 – 1486, 2009.

[92] Mahsa Shakeri, Hervé Lombaert, and Samuel Kadoury. *Machine Learning Meets Medical Imaging: First International Workshop, MLMMI 2015, Held in Conjunction with ICML 2015, Lille, France, July 11, 2015, Revised Selected Papers*, chapter Classification of Alzheimer's Disease Using Discriminant Manifolds of Hippocampus Shapes, pages 65–73. Springer International Publishing, Cham, 2015.

[93] I.A. Illn, J.M. Gorriz, J. Ramirez, D. Salas-Gonzalez, M.M. Lpez, F. Segovia, R. Chaves, M. Gmez-Rio, and C.G. Puntonet. 18f-fdg PET imaging analysis for computer aided alzheimers diagnosis. *Information Sciences*, 181(4):903 – 916, 2011.

[94] Igor Yakushev, Christian Landvogt, Hans-Georg Buchholz, Andreas Fellgiebel, Alexander Hammers, Armin Scheurich, Irene Schmidtmann, Alexander Gerhard, Mathias Schreckenberger, and Peter Bartenstein. Choice of reference area in studies of Alzheimer's disease using positron emission tomography with fluorodeoxyglucose-F18. *Psychiatry Research: Neuroimaging*, 164(2):143 – 153, 2008.

[95] Chong-Yaw Wee, Pew-Thian Yap, Daoqiang Zhang, Kevin Denny, Jeffrey N. Browndyke, Guy G. Potter, Kathleen A. Welsh-Bohmer, Lihong Wang, and Dinggang Shen. Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage*, 59(3):2045 – 2056, 2012.

[96] Edward Challis, Peter Hurley, Laura Serra, Marco Bozzali, Seb Oliver, and Mara Cercignani. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, 112:232 – 243, 2015.

[97] Biao Jie, Dinggang Shen, and Daoqiang Zhang. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II*, chapter Brain Connectivity Hyper-Network for MCI Classification, pages 724–732. Springer International Publishing, Cham, 2014.

[98] Y Zhang, N Schuff, G-H Jahng, W Bayne, S Mori, L Schad, S Mueller, A-T Du, JH Kramer, K Yaffe, et al. Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and Alzheimer disease. *Neurology*, 68(1):13–19, 2007.

[99] Alexander Hammers, Richard Allom, Matthias J Koepp, Samantha L Free, Ralph Myers, Louis Lemieux, Tejal N Mitchell, David J Brooks, and John S Duncan. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human brain mapping*, 19(4):224–247, 2003.

[100] Ioannis S. Gousias, Daniel Rueckert, Rolf A. Heckemann, Leigh E. Dyet, James P. Boardman, A. David Edwards, and Alexander Hammers. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2), 2008.

[101] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[102] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Jyrki Ltjnen, and Daniel Rueckert. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. *Medical Image Analysis*, 16(4):819 – 830, 2012.

[103] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Identifying informative imaging biomarkers via tree structured sparse learning for AD diagnosis. *Neuroinformatics*, 12(3):381–394, 2013.

[104] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, 220(2):841–859, 2015.

[105] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59 – 70, 2012.

[106] R. Chaves, J. Ramirez, J.M. Gorriz, M. Lpez, D. Salas-Gonzalez, I. lvarez, and F. Segovia. SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. *Neuroscience Letters*, 461(3):293 – 297, 2009.

[107] Jonathan Young, Alex Mendelson, M Jorge Cardoso, Marc Modat, John Ashburner, and Sebastien Ourselin. Improving MRI brain image classification with anatomical regional kernels. In *Machine Learning Meets Medical Imaging*, pages 45–53. Springer, 2015.

[108] V Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[109] Vladimir N Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[110] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, Mar 2002.

[111] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[112] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

[113] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006.

[114] Alan Julian Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, chapter Linear Discriminant Analysis, pages 237–280. Springer New York, New York, NY, 2008.

[115] Robin Wolz, Valtteri Julkunen, Juha Koikkalainen, Eini Niskanen, Dong Ping Zhang, Daniel Rueckert, Hilkka Soininen, Jyrki Lötjönen, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PloS one*, 6(10):e25446, 2011.

[116] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411, 2004.

[117] Linda K. McEvoy, Dominic Holland, Jr Donald J. Hagler, Christine Fennema-Notestine, James B. Brewer, and Anders M. Dale. Mild cognitive impairment: Baseline and longitudinal structural MR imaging measures improve predictive prognosis. *Radiology*, 259(3):834–843, 2011. PMID: 21471273.

[118] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967.

[119] Matthew D Mullin and Rahul Sukthankar. Complete cross-validation for nearest neighbor classifiers. In *ICML*, pages 639–646, 2000.

[120] Irina Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[121] J Ross Quinlan. *C4.5: Programs for machine learning*. Elsevier, 1993.

[122] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2 edition, 2009.

[123] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[124] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

[125] Lei Huang, Yaozong Gao, Yan Jin, Kim-Han Thung, and Dinggang Shen. Soft-split sparse regression based random forest for predicting future clinical scores of Alzheimers disease. In *Machine Learning in Medical Imaging*, pages 246–254. Springer, 2015.

[126] M. Liu, D. Zhang, E. Adeli, and D. Shen. Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, 63(7):1473–1482, July 2016.

[127] Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, Alzheimer's Disease Neuroimaging Initiative, et al. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuroimage*, 48(1):138–149, 2009.

[128] Xin Liu, Duygu Tosun, Michael W. Weiner, and Norbert Schuff. Locally linear embedding (LLE) for MRI based alzheimer's disease classification. *NeuroImage*, 83:148 – 157, 2013.

[129] Ramon Casanova, Fang-Chi Hsu, and for the Alzheimer's Disease Neuroimaging Initiative Mark A. Espeland. Classification of structural MRI images in Alzheimer's disease from the perspective of ill-posed problems. *PLoS ONE*, 7(10):1–12, 10 2012.

[130] B.S. Mahanand, S. Suresh, N. Sundararajan, and M. Aswatha Kumar. Identification of brain regions responsible for alzheimers disease using a self-adaptive resource allocation network. *Neural Networks*, 32:313 – 322, 2012. Selected Papers from IJCNN 2011.

[131] Saima Farhan, Muhammad Abuzar Fahiem, and Huma Tauseef. An ensemble-of-classifiers based approach for early diagnosis of Alzheimers disease: classification us-

ing structural features of brain images. *Computational and mathematical methods in medicine*, 2014, 2014.

[132] Sergey M Plis, Devon Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry Jeremy Bockholt, Jeffrey D Long, Hans J Johnson, Jane Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience*, 8(229), 2014.

[133] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[134] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[135] Mathias Fuchs and Norbert Krautenbacher. Minimization and estimation of the variance of prediction errors for cross-validation designs. *Journal of Statistical Theory and Practice*, 0(0):1–24, 0.

[136] Bradley T. Wyman, Danielle J. Harvey, Karen Crawford, Matt A. Bernstein, Owen Carmichael, Patricia E. Cole, Paul K. Crane, Charles DeCarli, Nick C. Fox, Jeffrey L. Gunter, Derek Hill, Ronald J. Killiany, Chahin Pachai, Adam J. Schwarz, Norbert Schuff, Matthew L. Senjem, Joyce Suhy, Paul M. Thompson, Michael Weiner, and Clifford R. Jack Jr. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & Dementia*, 9(3):332 – 337, 2013.

[137] Andrés Ortiz, Juan M Górriz, Javier Ramírez, Francisco Jesús Martínez-Murcia, Alzheimers Disease Neuroimaging Initiative, et al. LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimers disease. *Pattern Recognition Letters*, 34(14):1725–1733, 2013.

[138] Juergen Dukart, Karsten Mueller, Henryk Barthel, Arno Villringer, Osama Sabri, Matthias Leopold Schroeter, Alzheimer's Disease Neuroimaging Initiative, et al. Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI. *Psychiatry Research: Neuroimaging*, 212(3):230–236, 2013.

[139] Pierrick Coup, Simon F. Eskildsen, Jos V. Manjn, Vladimir S. Fonov, and D. Louis Collins. Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease. *NeuroImage*, 59(4):3736 – 3747, 2012.

[140] Zhengjia Dai, Chaogan Yan, Zhiqun Wang, Jinhui Wang, Mingrui Xia, Kuncheng Li, and Yong He. Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *Neuroimage*, 59(3):2187–2195, 2012.

[141] Katherine R Gray, Paul Aljabar, Rolf A Heckemann, Alexander Hammers, and Daniel Rueckert. Random forest-based manifold learning for classification of imaging data in dementia. In *Machine Learning in Medical Imaging*. Springer, 2011.

[142] Eric Westman, Andrew Simmons, J-Sebastian Muehlboeck, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Koszewska, Hilkka Soininen, Michael W. Weiner, Simon Lovestone, Christian Spenger, and Lars-Olof Wahlund. AddNeuroMed and ADNI: Similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. *NeuroImage*, 58(3):818 – 828, 2011.

[143] Gavin C. Cawley and Nicola L.C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, August 2010.

[144] Teddy Schall and Gary Smith. Do baseball players regress toward the mean? *The American Statistician*, 54(4):231–235, 2000.

[145] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

[146] R Bharat Rao, Glenn Fung, and Romer Rosales. On the dangers of cross-validation. An experimental evaluation. In *SIAM International Conference on Data Mining*, pages 588–596. Society for Industrial and Applied Mathematics, 2008.

[147] P.J Easterbrook, R Gopalan, J.A Berlin, and D.R Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867 – 872, 1991.

[148] Thomas Nowotny. Two challenges of correct validation in pattern recognition. *Frontiers in Robotics and AI*, 1(5), 2014.

[149] Alex F Mendelson, Maria A Zuluaga, Marco Lorenzi, Brian F Hutton, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Selection bias in the reported performances of AD classification pipelines. *NeuroImage: Clinical*, 14:400–416, 2017.

[150] Bruno Dubois, Howard H Feldman, Claudia Jacova, Steven T DeKosky, Pascale Barberger-Gateau, Jeffrey Cummings, Andr Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, Kenichi Meguro, John O'Brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Yaakov Stern, Pieter J Visser, and Philip Scheltens. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDSADRDA criteria. *The Lancet Neurology*, 6(8):734 – 746, 2007.

[151] Shiva Keihaninejad, Rolf A. Heckemann, Ioannis S. Gousias, Joseph V. Hajnal, John S. Duncan, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic MRI segmentation. *PLoS ONE*, 7(4):e33096, 04 2012.

[152] M Jorge Cardoso, Robin Wolz, Marc Modat, Nick C Fox, Daniel Rueckert, and Sebastien Ourselin. Geodesic information flows. In *Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS, pages 262–270. Springer, 2012.

[153] M Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C Fox, Sebastien Ourselin, Alzheimers Disease Neuroimaging Initiative, et al. STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical image analysis*, 17(6):671–684, 2013.

[154] Feng Shi, Bing Liu, Yuan Zhou, Chunshui Yu, and Tianzi Jiang. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus*, 19(11):1055–1064, 2009.

[155] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

[156] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991.

[157] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011.

[158] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11), 2009.

[159] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[160] Xinchuan Zeng and Tony R. Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12, 2000.

[161] Jose G Moreno-Torres, José A Sáez, and Francisco Herrera. Study on the impact of partition-induced dataset shift on-fold cross-validation. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(8):1304–1312, 2012.

[162] Qiong Zhang and Peter ZG Qian. Designs for crossvalidating approximation models. *Biometrika*, 100(4):997–1004, 2013.

[163] Maria A. Zuluaga, Ninon Burgos, Alex F. Mendelson, Andrew M. Taylor, and Sbastien Ourselin. Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries. *Medical Image Analysis*, 26(1):185 – 194, 2015.

[164] Ping Zhang. Model selection via multifold cross validation. *The Annals of Statistics*, 21(1):pp. 299–313, 1993.

[165] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 1997.

[166] Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

[167] Mathias Fuchs, Roman Hornung, Riccardo De Bin, and Anne-Laure Boulesteix. A U-statistic estimator for the variance of resampling-based error estimators. *arXiv preprint arXiv:1310.8203*, 2013.

[168] Gavin C. Cawley and Nicola L.C. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585 – 2592, 2003.

[169] Gavin C Cawley and Nicola LC Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural networks*, 17(10):1467–1475, 2004.

[170] Douglas R Stinson. *Combinatorial designs: constructions and analysis*. Springer Science & Business Media, 2007.

[171] Mathias Fuchs and Norbert Krautenbacher. A variance decomposition and a central limit theorem for empirical losses associated with resampling designs. 2014.

[172] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.

[173] Edward R Dougherty, Amin Zollanvari, and Ulisses M Braga-Neto. The illusion of distribution-free small-sample classification in genomics. *Current genomics*, 12(5):333–341, 2011.

[174] Daniel Berrar and Jose A. Lozano. Significance tests or confidence intervals: which are preferable for the comparison of classifiers? *Journal of Experimental & Theoretical Artificial Intelligence*, 25(2):189–206, 2013.

[175] Jiming Jiang. *Large sample techniques for statistics*. Springer Science & Business Media, 2010.

[176] Vidmantas Bentkus and F Gotze. The berry-esseen bound for student's statistic. *The Annals of Probability*, pages 491–503, 1996.

[177] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 2001.

[178] Yves Grandvalet and Yoshua Bengio. Hypothesis Testing for Cross-Validation. *Montreal Universite de Montreal, Operationnelle DdIeR*, 2006.

[179] Remco R Bouckaert. Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003.

[180] Edwin Hewitt and Leonard J Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 1955.

[181] Ethem Alpaydin. Combined $5\times$ 2 cv F test for comparing supervised classification learning algorithms. *Neural computation*, 11(8):1885–1892, 1999.

[182] Marianthi Markatou, Hong Tian, Shameek Biswas, and George M Hripcsak. Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, 6:1127–1168, 2005.

[183] Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985.

[184] Remco R Bouckaert. Choosing learning algorithms using sign tests with high replicability. In *AI 2003: Advances in Artificial Intelligence*. Springer, 2003.

[185] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining*. Springer, 2004.

[186] Remco R Bouckaert. Low replicability of machine learning experiments is not a small data set phenomenon. *ICML Meta Learning workshop*, 2005.

[187] Remco R Bouckaert. Estimating replicability of classifier learning experiments. In *Proceedings of the twenty-first international conference on Machine learning*, page 15. ACM, 2004.

[188] Qing Wang and Bruce Lindsay. Variance estimation of a general U-statistic with application to cross-validation. *Statistica Sinica*, 24:1117–1141, 2014.

[189] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 1948.

[190] Polina Golland and Bruce Fischl. Permutation tests for classification: Towards statistical significance in image-based studies. In *Information processing in medical imaging*. Springer, 2003.

[191] Phillip I Good. *Permutation, parametric and bootstrap tests of hypotheses*, volume 3. Springer, 2005.

[192] Bo Cheng, Daoqiang Zhang, and Dinggang Shen. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part I*, chapter Domain Transfer Learning for

MCI Conversion Prediction, pages 82–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[193] L. Khedher, J. Ramirez, J.M. Gorriz, A. Brahim, and F. Segovia. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, 151, Part 1:139 – 150, 2015.

[194] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, 60(2), 2012.

[195] Kristin A. Linn, Bilwaj Gaonkar, Theodore D. Satterthwaite, Jimit Doshi, Christos Davatzikos, and Russell T. Shinohara. Control-group feature normalization for multivariate pattern analysis of structural MRI data using the support vector machine. *NeuroImage*, 132:157 – 166, 2016.

[196] Alex F Mendelson, Maria A Zuluaga, Brian F Hutton, and Sebastien Ourselin. Bolstering heuristics for statistical validation of prediction algorithms. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, pages 77–80. IEEE, 2015.

[197] Morton B Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992, 1975.

[198] James T Kost and Michael P McDermott. Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190, 2002.

[199] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[200] Vladimir Vovk. Combining p-values via averaging. *arXiv preprint arXiv:1212.4966*, 2012.

[201] John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.

[202] Xiaojin Zhu. Semi-supervised learning. In *Encyclopedia of Machine Learning*, pages 892–897. Springer, 2011.

[203] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[204] Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for K-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*. ACM, 1999.